LEARNING REUSABLE INFERENCE ACROSS REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Good in-context prediction does not establish whether a model has learned inference that transfers across representations. We construct a controlled test using latent parity tasks whose full-rank binary embeddings change representation reuse while preserving the Bayes problem and the statistical value of joint evidence. An exact hierarchy of support-label interactions quantifies this value: at latent rank r, Bayes prediction needs $O(r)$ supports and polynomial-time inference, whereas the optimal label-additive predictor needs $\Theta(2^r)$ supports at fixed risk. Across two contextual architectures, final-answer training exploits more of this opportunity with persistent embeddings, but that benefit transfers incompletely to unseen embeddings. Controls show that joint-label dependence is learnable when the organisation of the evidence is supplied; difficulty at higher ranks even with persistent embeddings motivates supervision of the inference process. A process-supervised recurrent learner executes autonomously under fresh embeddings, transfers beyond its training ranks, and turns additional supports into lower risk. These results separate statistical opportunity, representation reuse, and acquired computation, showing why evaluations of learned adaptation should test both what transfers and how effectively additional context is used.

1 INTRODUCTION

In-context learning (ICL) allows a model to adapt its predictions from labelled examples without updating its parameters. Dataset-conditioned models such as Prior-Data Fitted Networks (PFNs), TabPFN, and TabICL make this explicit: an entire dataset becomes the input to a learned predictor (Müller et al., 2022; Hollmann et al., 2025; Qu et al., 2025). Training across tasks can amortise the construction of an inference procedure, but its usefulness on a new task depends on what it extracts from that task's support set. Amortisation concerns work shared across tasks; within-task sample efficiency concerns how many supports adaptation requires.

Good in-context prediction alone does not establish how broadly that inference procedure transfers. A model may genuinely infer a new rule while relying on representation-specific structure shared across training tasks. We therefore ask: *what part of learned adaptation remains useful when the representation changes?* Answering this requires a control: changing representation must not also reduce the predictive information in context.

We construct such a control using latent parity tasks. Our construction varies whether a full-rank binary embedding persists across episodes or is sampled afresh, while preserving the Bayes prediction problem and the statistical value of combining support labels. Consequently, a performance difference measures a change in the learner's ability to exploit the evidence, rather than a change in what that evidence makes possible (Figure 1).

To quantify this opportunity, we organise predictors by how many support labels enter each term jointly, allowing every term to use all inputs. The optimal risks in this hierarchy benchmark the benefit unavailable without joint label dependence. In the latent family, Bayes prediction needs only $O(r)$ supports at rank r, whereas even the best label-additive predictor needs $\Theta(2^r)$ at fixed risk. This advantage coexists with polynomial-time inference.

The experiments then ask when learners recover that advantage. Endpoint training benefits from persistent representations, but the benefit transfers incompletely to unseen embeddings. Controls

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.

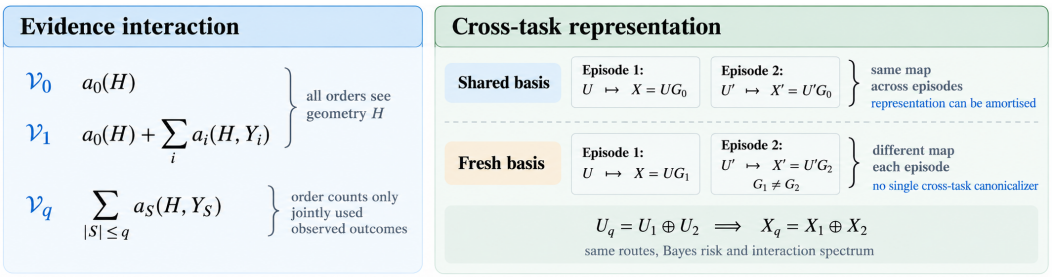


Figure 1: **Within-task evidence and cross-task representation.** $\mathcal{V}_q$ restricts the number of support labels entering each term jointly; every term may use all support and query inputs H. Full-rank embeddings preserve the Bayes-relevant linear relations and interaction spectrum, while fixed versus fresh embeddings change the representation available for reuse across episodes.

show that the learners can exploit joint labels when their useful organisation is supplied, motivating supervision of the inference process itself. A process-supervised recurrent learner executes autonomously under fresh embeddings, transfers to harder ranks, and makes productive use of more evidence. Our contributions are:

- **A representation intervention with exact statistical controls.** We prove embedding invariance of Bayes risk and every interaction-order risk, separating representation transfer from loss of predictive information (Propositions 4–5).
- **An explicit support-sample benchmark.** Applying classical projection and character orthogonality gives $R(\mathcal{V}_q) = \Pr(k_{\min} > q)$, where $k_{\min}$ is the smallest number of support labels needed to determine the query. This yields $O(r)$ Bayes versus $\Theta(2^r)$ optimal additive support requirements, independently of optimisation (Theorem 6).
- **Tests of representation reuse and learned execution.** We test which benefits of persistent embeddings transfer, and whether process supervision teaches computation that uses more evidence at unseen ranks (Sections 5–6).

2 Related work

Dataset-conditioned prediction. Bayesian accounts connect adaptation to task inference (Grant et al., 2018; Xie et al., 2022; Panwar et al., 2024); PFNs amortise posterior prediction from datasets (Müller et al., 2022), with TabPFN and TabICL targeting tabular prediction (Hollmann et al., 2023; 2025; Qu et al., 2025). Nagler (2023) analyses behaviour as inference-time data grow. We resolve that statistical opportunity by the number of support labels that must interact, then preserve it across representation changes. This connects the value of a larger context to a test of whether learned inference continues to extract it.

Parity and representation transfer. Bhattamishra et al. (2024) study discrete-function ICL and teaching sequences; Park et al. (2024) report architecture-dependent sparse-parity performance. Abedsoltan et al. (2025) study generalisation to held-out coordinate subsets with chain-of-thought supervision: their exponential quantity is the number of tasks covered by relatively few training tasks. Our exponential separation concerns supports within a new episode. Shi et al. (2024) use an unknown shared real-valued dictionary with two-feature parity and informative feature–label correlations. Recent linear-regression analyses study ICL under subspace or task-distribution shift (Kwon et al., 2026; Letey et al., 2026). Our binary embeddings instead vary across episodes while preserving the clean-parity Bayes problem and its interaction-risk curve, making representation reuse testable at fixed statistical opportunity.

Process supervision and learned algorithms. Kim & Suzuki (2025) prove optimisation benefits of intermediate parity supervision for a predictor trained from sample-averaged gradients. We test whether supervising relations among a new task's supports teaches contextual computation that transfers in rank and uses more evidence. This links intermediate supervision to an exact contextual

sample-efficiency benchmark, complementing compositional and length-generalisation studies (Wen et al., 2025; Zhou et al., 2024; Veličković & Blundell, 2021).

Classical orthogonality, a different restriction. Parity orthogonality and statistical-query (SQ) lower bounds are classical (O'Donnell, 2014; Kearns, 1998). Blum et al. (2003) also study queries about tuples of examples. SQ bounds restrict access to approximate population statistics; $\mathcal{V}_q$ restricts the final predictor's dependence on observed support labels while allowing arbitrary use of their inputs. Our finite-support risk identity therefore describes a different restriction, with the additional control that its entire risk curve is unchanged by the embedding intervention.

Interaction spaces and inference mechanisms. Nested L^2 projections, functional analysis of variance (ANOVA), and Boolean Fourier analysis provide the mathematical tools (Hoeffding, 1948; Hooker, 2007; Chastaing et al., 2012; O'Donnell, 2014). We apply them to the posterior predictive mean as a function of support labels; the projection identity itself is standard. Deep Sets and Set Transformers can create label interactions through nonlinear aggregation (Zaheer et al., 2017; Lee et al., 2019), although this expressivity does not establish equal algorithmic capacity for particular finite architectures. Likewise, ICL constructions implementing gradient descent (Akyürek et al., 2023; Von Oswald et al., 2023; Gatmiry et al., 2024) do not imply that all contextual inference is label-additive. Our distinction also complements task recognition versus task learning (Pan et al., 2023): inferring a new rule can still rely on a representation shared across training tasks.

3 Statistical opportunity from joint evidence

We first measure what useful evidence is available. Standard Hilbert-space projections quantify the value of combining labels while giving the predictor unrestricted access to all inputs.

3.1 A support-outcome hierarchy

Fix a support size n and episode distribution Π. Let $H = (X_{1:n}, X_{\mathrm{q}})$, $Y = Y_{1:n}$, and assume $\mathbb{E}Y_{\mathrm{q}}^2 < \infty$. Under squared loss, define

$$m_\Pi(H,Y) = \mathbb{E}[Y_{\mathrm{q}} \mid H, Y], \qquad R(g) = \mathbb{E}(Y_{\mathrm{q}} - g(H,Y))^2, \qquad R^\star = R(m_\Pi). \tag{1}$$

For $S \subseteq [n]$, L_S^2 contains functions using all of H and only the labels indexed by S, with finite expected square under Π.

Definition 1 (Support-outcome interaction spaces)**.** For $q = 0, \dots, n$, let

$$\mathcal{V}_q = \overline{\mathrm{span}}\left(\bigcup_{S \subseteq [n]:\, |S| \le q} L_S^2 \right), \qquad \mathcal{V}_0 \subseteq \mathcal{V}_1 \subseteq \cdots \subseteq \mathcal{V}_n, \tag{2}$$

and let $P_q f$ be the closest element of $\mathcal{V}_q$ to f in expected squared distance. The overline includes limits in this distance.

Thus $\mathcal{V}_1$ contains sums $a_0(H) + \sum_i a_i(H, Y_i)$; for sign labels these are exactly $a_0(H) + \sum_i b_i(H) Y_i$ up to closure. The restriction concerns labels entering a term jointly, not how many labels the predictor uses overall. For example, ridge regression gives

$$\widehat{y}_{\mathrm{q}} = \phi(X_{\mathrm{q}})^\top (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top Y, \qquad \Phi_{i,:} = \phi(X_i)^\top.$$

With fixed features and label-independent $\lambda > 0$, this is linear in Y even when ϕ is nonlinear in the inputs or learned across tasks. It therefore lies in $\mathcal{V}_1$, as do nearest-neighbour regression and input-weighted neighbour averages. Nonlinearity in the inputs is distinct from joint dependence on the episode's labels.

Call the sequence of projection energies the interaction spectrum, and define its tail by

$$I_0 = \|P_0 m_\Pi\|_2^2, \qquad I_q = \|(P_q - P_{q-1}) m_\Pi\|_2^2, \qquad \Gamma_{>q} = \|m_\Pi - P_q m_\Pi\|_2^2 = \sum_{j>q} I_j. \tag{3}$$

Nested orthogonal projections give this identity even for dependent labels. Here $P_0 m_\Pi = \mathbb{E}[Y_{\mathrm{q}} \mid H]$, so I_0 measures prediction from inputs alone.

Theorem 2 (Bayes interaction risk identity). *For every $q \in \{0, \dots, n\}$,*

$$R(\mathcal{V}_q) - R^\star := \inf_{g \in \mathcal{V}_q} R(g) - R^\star = \Gamma_{>q}. \tag{4}$$

Conditional expectation gives $R(g) - R^\star = \|g - m_\Pi\|_2^2$; minimising this distance over $\mathcal{V}_q$ yields $P_q m_\Pi$. Thus $\Gamma_{>q}$ is the unavoidable excess risk of every order-q predictor, even with oracle coefficient functions. Risk below $R(\mathcal{V}_1)$ certifies useful nonadditive evidence use, but scalar risk does not recover a model's interaction spectrum or identify its internal algorithm. The benchmark quantifies opportunity rather than learnability.

3.2 A CONCRETE ADDITIVE ADAPTATION MECHANISM

The ridge example extends to a familiar adaptation procedure. Suppose the support objective is quadratic in a prediction state w, with

$$\nabla_w L_H(w; Y) = A_H w - B_H Y - c_H, \qquad w_{t+1} = w_t - M_{t,H} \nabla_w L_H(w_t; Y). \tag{5}$$

The initial state, coefficients, preconditioners, and number of updates are label-independent, and the query head is $b_H + z_H^\top w_K$.

Theorem 3 (Fixed-geometry quadratic adaptation). *For any finite K, $w_K = \alpha_K(H) + C_K(H)Y$. Consequently, every square-integrable query predictor of this form lies in $\mathcal{V}_1$ and has excess risk at least $\Gamma_{>1}$.*

The affine form is preserved by the recurrence in (5); Appendix C gives the proof and an approximate-class extension. Features learned across tasks or sampled independently of episode labels are covered. More features or updates preserve the affine dependence on Y under these assumptions.

Updating a multilayer perceptron's (MLP's) hidden layers on the episode can create nonlinear label dependence, whether initialisation is meta-learned or fresh. The boundary is the complete adaptation-and-prediction map. Even one additive update can produce interactions: if $w^+ = a + \sum_i b_i Y_i$, the head $(w^+)^2$ contains terms $2b_i b_j Y_i Y_j$. Instance-wise support losses therefore do not imply an additive final prediction. Nonlinear model-agnostic meta-learning (MAML) and Meta-SGD can also escape $\mathcal{V}_1$ (Finn & Levine, 2018; Collins et al., 2022). Our MAML and Meta-SGD baselines adapt all predictor parameters and thus fall outside the theorem's guarantee. The theorem characterises a concrete additive mechanism, while the experiments test what richer learners acquire.

4 REPRESENTATION-RANDOMISED LATENT CODES

We embed clean parity in a family that varies representation reuse while preserving the statistical opportunity quantified above. Draw a task $z \sim \mathrm{Unif}(\mathbb{F}_2^r)$ and independent uniform inputs $U_1, \dots, U_n, U_\mathrm{q} \in \mathbb{F}_2^r$, with $Y_i = (-1)^{\langle z, U_i \rangle}$. Expose $X_i = U_i G$ and $X_\mathrm{q} = U_\mathrm{q} G$ for a full-row-rank $G \in \mathbb{F}_2^{r \times D}$, independent of the task code and latent inputs. The embedding may be fixed across episodes or sampled afresh. Whenever $X_\mathrm{q} = \bigoplus_{i \in S} X_i$, the query label is $Y_\mathrm{q} = \prod_{i \in S} Y_i$. Define

$$k_{\min}(H) = \min\{|S| : X_\mathrm{q} = \textstyle\bigoplus_{i \in S} X_i\}, \tag{6}$$

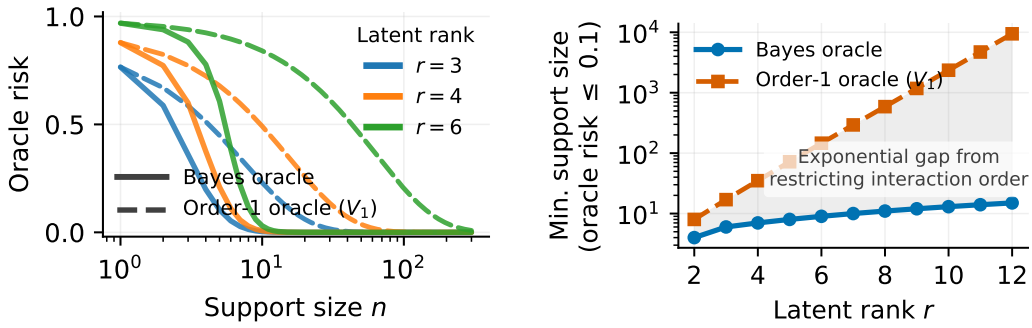
with $k_{\min} = \infty$ outside the support span and $k_{\min} = 0$ for a zero query. A *route* is any subset satisfying this relation; it need not be a minimum-cardinality subset.

4.1 REPRESENTATION INVARIANCE AND SHARED COORDINATES

Proposition 4 (Representation invariance of statistical opportunity). *Fix r, n, D with $D \geq r$ and the uniform latent-code law above. For any full-row-rank G, the coupled latent and observed geometries have identical route sets:*

$$\bigoplus_i c_i U_i = U_\mathrm{q} \quad \Longleftrightarrow \quad \bigoplus_i c_i X_i = X_\mathrm{q}, \qquad c \in \mathbb{F}_2^n. \tag{7}$$

They therefore have the same span membership, $k_{\min}$, conditional Bayes prediction and conditional support-outcome projection energies. Consequently $R^\star$, every I_q and $\Gamma_{>q}$, and every fixed-risk support threshold n_{Bayes} and $n_{\mathcal{V}_q}$ are unchanged by choosing a fixed embedding or an independent random embedding each episode.



(a) Oracle risk versus support size.

(b) Supports for squared risk at most 0.1.

Figure 2: **Joint evidence creates a large statistical opportunity.** The latent-code Bayes and label-additive oracles have $O(r)$ and $\Theta(2^r)$ fixed-risk support requirements. These exact statistical quantities are invariant to the full-rank embedding and independent of a learned model.

Proposition 4 supplies the control needed to interpret the experiment: the representation changes while the oracle value of the evidence stays fixed. Injectivity gives (7). The conditional label law is uniform on the binary linear relations determined by the geometry; the same relations give the same conditional projection problem. These prediction problems agree under a coupling, although their marginal input laws differ (Appendix D.1).

Proposition 5 (Shared canonicalisation and coordinate ambiguity)**.** *A fixed embedding G_0 admits a right inverse R_0 with $G_0R_0 = I_r$, so $X_iR_0 = U_i$ across episodes. For $r \geq 2$, if the embedding law includes both G and AG for some $A \in GL(r, 2)$ with $A \neq I_r$, no single linear map recovers the sampled latent coordinates for both. Moreover, the transformations*

$$U_i' = U_iA, \qquad G' = A^{-1}G, \qquad z' = zA^{-T} \tag{8}$$

leave every observed input and label unchanged (using row-vector task codes). Thus the true latent coordinates are not identifiable from an episode.

Prediction can instead reuse observable relations: if $X_q = X_2 \oplus X_5$, then $Y_q = Y_2Y_5$ under every full-rank embedding. In general the relevant object is the relation class $\mathcal{C}(H) = \{c \in \mathbb{F}_2^n : \bigoplus_i c_iX_i = X_q\}$. For any $c \in \mathcal{C}(H)$, prediction is $\prod_i Y_i^{c_i}$. Thus prediction can use observable relations without recovering the sampled latent coordinates. Under the uniform latent law, embeddings with the same row space induce identical observed episode laws: replacing G by AG simply relabels uniform U and z. The experimental ladder varies the observed subspace as well as its embedding; it does not test an unobservable basis change within one subspace.

4.2 Exact spectrum and efficient inference

Theorem 6 (Latent-code spectrum and sample efficiency)**.** *The Bayes predictor is the label product along any valid route, or zero outside the support span. Moreover,*

$$\begin{aligned} R^\star &= \Pr(k_{\min} = \infty), \qquad I_q = \Pr(k_{\min} = q), \\ R(\mathcal{V}_q) &= \Pr(k_{\min} > q), \qquad \Gamma_{>q} = \Pr(q < k_{\min} < \infty). \end{aligned} \tag{9}$$

In particular,

$$R^\star \leq 2^{r-n}, \qquad R_{\mathcal{V}_1} = (1 - 2^{-r})^{n+1}. \tag{10}$$

For fixed risk $\epsilon \in (0, 1)$, Bayes needs $O(r)$ supports, whereas $\mathcal{V}_1$ needs $\Theta(2^r)$. More generally the respective upper bound and additive rate are $O(r + \log(1/\epsilon))$ and $\Theta(2^r \log(1/\epsilon))$ in the small-risk regime.

Conditioned on H, support-label products are characters of the uniform task code. Distinct characters are orthogonal. A term involving at most q labels can contain the query character exactly when $k_{\min} \leq q$, proving the spectrum. The additive oracle recognises zero queries and individual support matches, which become rare as rank grows. Joint inference can instead combine supports to determine

Table 1: **Statistical benchmarks across task families.** $\epsilon \in (0, 1)$ is fixed; $d = \rho^{-2}$ grows in distributed evidence. The complete-factor control supplies exactly k labels, so it has no support-size scaling experiment.

Task	Bayes	Optimal label-additive
Complete factors ($n = k \geq 2$)	Risk 0	Risk 1
Distributed evidence	$n_\epsilon = \Theta(d)$	$n_\epsilon = \Theta(d)$
Latent codes	$n_\epsilon = \Theta(r)$	$n_\epsilon = \Theta(2^r)$

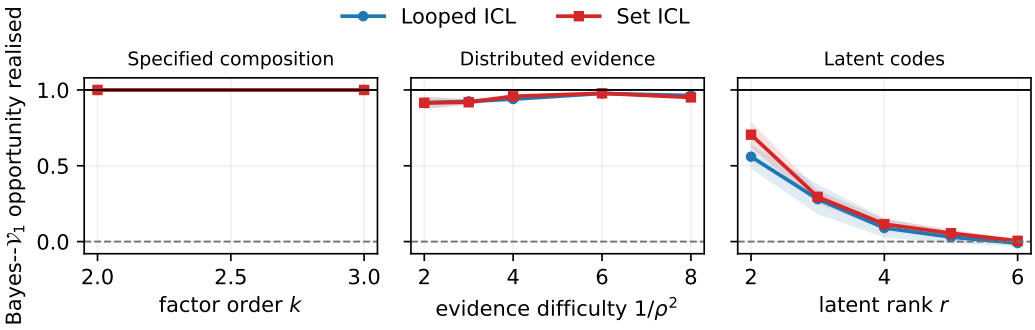


Figure 3: **Joint evidence use is learnable when its organisation is supplied.** Looped ICL and Set ICL recover nearly all Bayes–$\mathcal{V}_1$ opportunity on complete-factor composition and distributed evidence, but little at larger latent ranks. The controls supply all factors to multiply or a common aggregate; latent codes require finding which labels to combine. These are different task families, not a matched intervention on interaction order alone. Bands are pointwise 95% intervals across five seeds.

an unseen query. A rank-deficiency bound gives the Bayes rate (Figure 2). Complete proofs appear in Appendix D.

This applies classical character orthogonality (O'Donnell, 2014) to support labels rather than input coordinates: $k_{\min}$ determines when the query character first appears. The resulting risk curve is preserved across embeddings. Even oracle optimisation within $\mathcal{V}_1$ cannot remove its support requirement.

For $n = \lceil \alpha r \rceil$, $\alpha > 1$, the required order concentrates around $nh_2^{-1}(1/\alpha)$, where h_2 is binary entropy on $[0, 1/2]$; almost all Bayes target energy therefore lies at orders $\Theta(r)$. Nevertheless, Gaussian elimination finds a valid route in polynomial time; prediction does not require a minimum-size route. The separation is in support-sample complexity, not computational hardness. Appendix E also constructs recurrent hard-attention inference in $r + O(1)$ rounds, showing that high-order evidence use admits compact contextual computation under its stated assumptions. Whether the experimental learners acquire such computation remains an empirical question.

5 What makes the opportunity learnable?

We compare Looped ICL, which repeatedly updates query/workspace tokens using support memory, and Set ICL, which jointly processes support and query tokens. Both are trained on final-answer squared error. We report the fraction of Bayes–additive opportunity recovered,

$$\eta = \frac{R_{\mathcal{V}_1} - R(F_\phi)}{R_{\mathcal{V}_1} - R^\star}, \tag{11}$$

when the denominator is positive. Bayes corresponds to one, the additive oracle to zero, and negative values mean worse-than-additive risk. Means and pointwise 95% Student-t intervals use five training seeds; pairwise comparisons are descriptive. Validation risk selects models, excluding held-out ranks. Appendix A gives protocols, hyperparameters, and uncertainty definitions.

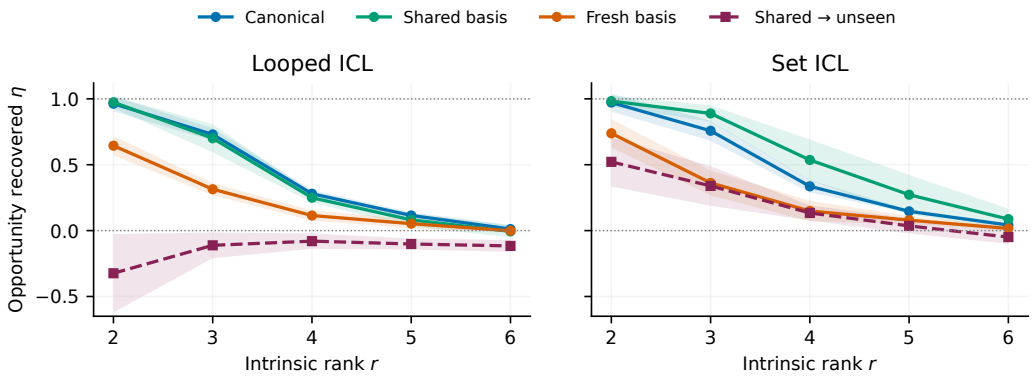


Figure 4: **Persistent representations help exploit a fixed opportunity, but the benefit transfers incompletely.** The same ladder for Looped ICL (left) and Set ICL (right). Persistent embeddings improve mean recovery at lower ranks, but the shared-to-unseen evaluation (dashed) exposes limited transfer. Fresh-embedding training and unseen-embedding evaluation test acquisition and transfer respectively. Bands are pointwise 95% intervals across five seeds; fine pairwise orderings are descriptive.

5.1 Interaction order alone does not determine endpoint difficulty

Can a learner combine labels when the useful combination is supplied, and can it discover which labels to combine when it is not? We begin with controls that supply the organisation of the evidence, before testing latent codes where a useful relation must be found from the inputs.

Fixed composition and a shared aggregate. The factor controls ask for $\prod_{j=1}^{k} Y_j$ with all $k \in \{2, 3\}$ required factors supplied and no distractors. Distributed evidence instead predicts a uniform sign Z from $Y_i = Z\xi_i$, where the independent noise signs have mean $\rho \in (0, 1)$. For $S_n = \sum_i Y_i$,

$$m_{\mathrm{DE}}(Y) = \tanh\big(\operatorname{atanh}(\rho) S_n\big). \tag{12}$$

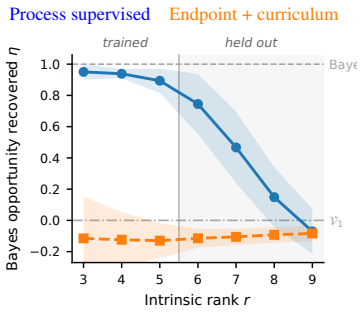
For $n \geq 3$, a nonlinear function of one scalar thus realises a nonadditive Bayes rule. Table 1 distinguishes the oracle comparisons; the controls are not matched interventions on interaction order. Appendix A.1 gives their exact oracles.

Both learners recover nearly all oracle opportunity on composition and distributed evidence, but little by latent rank six (Figure 3). Thus their latent-code failure cannot be described simply as an inability to represent or learn any joint label dependence. What the controls supply, and latent codes require the learner to find, is a useful organisation of the evidence. Route-revealed interventions recover higher-order prediction, while weaker structural losses give partial gains and a matched random cue does not reproduce the benefit (Appendix B.3). This supports routing as an obstacle without establishing it as the sole cause. MAML and Meta-SGD results remain supplementary: their failures on order-three composition make latent-code failure less diagnostic of routing (Appendix B).

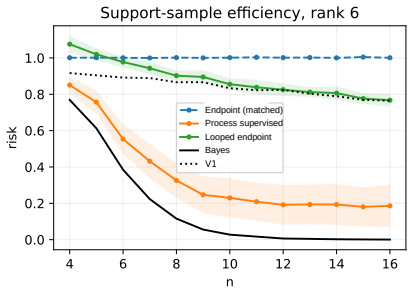
5.2 Representation alignment as an amortisable resource

We now vary representation reuse while preserving the latent task and support sizes. Proposition 4 guarantees that the entire oracle risk curve in (9) is unchanged. Three training conditions use canonical coordinates (A), one shared random embedding (B), or a fresh embedding per episode (C). A fourth evaluation applies B's trained model to previously unseen shared embeddings (D); D is a representation shift, not a fourth training distribution. All ranks 2–6 participate in training and validation. This experiment tests representation generalisation within the training rank range, separately from the held-out-rank experiment below.

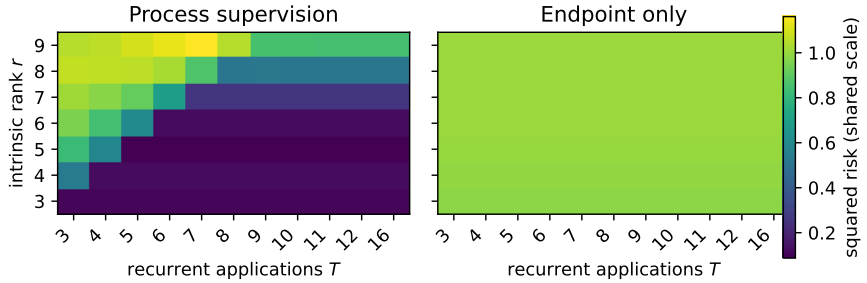
Figure 4 supports a specific distinction: *learning new rules within a persistent embedding does not ensure transfer of the inference procedure to new embeddings*. At lower ranks, canonical and shared-random training both recover more opportunity than fresh-embedding training in the seed means. The



(a) Rank transfer, $T = r$ (width 64).



(b) Support efficiency at $r = 6$ (width 64).



(c) Depth scaling (matched width 256).

Figure 5: **Process supervision teaches computation that transfers and uses additional evidence.** (a) A process-supervised transition transfers beyond the training ranks 3–5, unlike the stronger endpoint-only curriculum control. (b) At held-out $r = 6$, process-supervised risk falls as supports increase; the matched endpoint and Looped endpoint references do not approach Bayes. (c) On one shared risk scale, repeated process-supervised transitions become useful as depth reaches the required rank; additional endpoint depth does not recover this behaviour. Panels use the indicated protocols and widths, not a pooled estimate; (a)'s curriculum control differs from (b,c)'s matched endpoint controls. Bands summarise five-seed uncertainty.

shared condition matters: the advantage is not confined to canonical coordinates. For Looped ICL, shared-to-unseen evaluation falls below both its shared-basis performance and fresh-basis training. For Set ICL, the unseen and fresh curves are closer at several ranks. The dashed curve tests transfer after shared training; the fresh curve tests acquisition under representation variation. Similar risks need not imply similar mechanisms.

The oracle invariance rules out less informative evidence as an explanation. Persistent observed subspaces offer reusable structure, but these results do not identify a learned inverse or exclude other input-dependent shortcuts. At rank six, even aligned conditions recover little opportunity. Persistent structure therefore helps without resolving the acquisition problem, motivating the next test: whether exposing the inference process can teach computation that transfers as the intrinsic problem becomes harder.

6 Acquiring Reusable Computation Through Process Supervision

The representation ladder tests whether inference survives a change in observed inputs; rank transfer asks whether it remains useful as the underlying problem becomes harder. We connect these tests to the known benefits of intermediate parity supervision (Kim & Suzuki, 2025) by asking whether a learned process can exploit the statistical value of context under fresh embeddings and at higher ranks.

A separate recurrent Set Transformer is trained to predict elimination transitions. At evaluation it predicts all states and answers autonomously, without oracle states, pivots, span tests, or routes. Fresh embeddings are used throughout; training and validation use ranks 3–5, with ranks 6–9 held

out at fixed ambient dimension $D = 12$. Test support lengths remain within the training range, isolating transfer in intrinsic complexity from extrapolation in input dimension or context length. Appendix A.6 specifies the state, losses, and teacher-forcing schedule.

Figure 5 connects execution to statistical benefit. At depth $T = r$, the learned transition recovers substantial opportunity at unseen ranks six and seven, with declining recovery thereafter. The endpoint-only curriculum control remains near or below the additive reference. At held-out rank six, adding supports lowers process-supervised risk well below that reference, whereas the matched endpoint model remains near uninformative risk. The model therefore uses joint evidence to realise part of the known support-sample advantage; the support sweep thus tests evidence use beyond fixed-context accuracy, without establishing an asymptotic $O(r)$ rate for the learned model. Finally, the depth sweep shows that additional recurrent computation becomes useful after process supervision. More endpoint-only depth does not produce the same behaviour under these budgets.

The matched endpoint control shares the mutable-state architecture and readout. Looped ICL instead updates query/workspace tokens against fixed support memory, so its comparison is not an architecture-controlled test of supervision. State-only supervision and removing teacher forcing give smaller gains; pivot-only supervision does not reproduce successful execution. Continuation after removing auxiliary supervision also loses higher-order recovery, while retaining much order-one retrieval (Appendix B.2). Autonomous execution, acquisition from final answers, and retention during further training are therefore separate empirical requirements. We show gradient diagnostics and a restricted analytic example appear in Appendices B.4 and F; they do not establish the cause of endpoint failure.

7 Discussion

Good in-context prediction is not enough to establish reusable inference. Our framework makes this distinction testable by first quantifying what the context permits. The support-outcome hierarchy measures the predictive value of combining labels; in latent codes, that value yields an exponential support-sample advantage over the optimal additive predictor. Because polynomial-time inference can realise the advantage, high interaction order need not imply computational intractability.

The representation intervention then holds this opportunity fixed while varying what structure can be reused across tasks. The tested endpoint learners benefit from persistent embeddings, yet transfer that benefit incompletely to unseen embeddings. They can therefore learn new task rules without acquiring equally transferable inference. A representation-specific procedure can still perform sophisticated adaptation; the point is that fixed-representation task generalisation leaves its transfer unresolved. Evaluations of learned inference should distinguish these capabilities and measure performance against the information available in context.

Algorithm-aligned process supervision provides a constructive counterpart. Exposing elimination transitions during training produces autonomous computation that remains useful under fresh embeddings and at some higher ranks. Its gains from additional supports and recurrent depth connect that transfer to the statistical opportunity. The finite-rank results realise part of the oracle advantage, while declining recovery and losses after auxiliary supervision is removed show that acquisition, execution, and retention remain distinct requirements.

Extending these distinctions beyond the algebraic setting requires replacing exact controls by empirical approximations: estimating the support-outcome interaction spectrum, and constructing representation changes that preserve the statistical value of context while altering what can be reused across tasks. These questions are especially natural for dataset-conditioned predictors such as PFNs and tabular foundation models (Müller et al., 2022; Qu et al., 2025), whose predictive performance alone does not reveal how support-label evidence is integrated or which parts of the learned procedure transfer across representations. Our controlled setting makes these quantities separately identifiable, providing an exact reference point for studying evidence use and reusable inference in less controlled domains.

AI USE STATEMENT

LLMs assisted with literature discovery, manuscript drafting and editing. The authors are responsible for the accuracy of the claims, citations, and final submission.

REPRODUCIBILITY STATEMENT

Appendix A provides task definitions and sampling procedures, model architectures, training budgets, hyperparameters, checkpoint selection, random seeds, evaluation protocols, and uncertainty summaries. The remaining appendices provide the assumptions and proofs for the theoretical results.

REFERENCES

Amirhesam Abedsoltan, Huaqing Zhang, Kaiyue Wen, Hongzhou Lin, Jingzhao Zhang, and Mikhail Belkin. Task generalization with autoregressive compositional structure: Can learning from D tasks generalize to D^T tasks? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 154–173. PMLR, 2025. URL https://proceedings.mlr.press/v267/abedsoltan25a.html.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Satwik Bhattamishra, Arkil Patel, Phil Blunsom, and Varun Kanade. Understanding in-context learning in transformers and LLMs by learning to learn discrete functions. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f7e7fabd73b3df96c54a320862afcb78-Abstract-Conference.html.

Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM*, 50(4):506–519, 2003. doi: 10.1145/792538.792543. URL https://arxiv.org/abs/cs/0010022.

Gaëlle Chastaing, Fabrice Gamboa, and Clémentine Prieur. Generalized Hoeffding–Sobol decomposition for dependent variables—application to sensitivity analysis. *Electronic Journal of Statistics*, 6:2420–2448, 2012. doi: 10.1214/12-EJS749.

Gérard Cohen, Iiro Honkala, Simon Litsyn, and Antoine Lobstein. *Covering Codes*. North-Holland, Amsterdam, 1997. ISBN 9780444825117.

Liam Collins, Aryan Mokhtari, Sewoong Oh, and Sanjay Shakkottai. MAML and ANIL provably learn representations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4238–4310. PMLR, 2022. URL https://proceedings.mlr.press/v162/collins22a.html.

Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269–281, 1979. doi: 10.1214/aos/1176344611.

Chelsea Finn and Sergey Levine. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HyjC5yWCW.

Khashayar Gatmiry, Nikunj Saunshi, Sashank J. Reddi, Stefanie Jegelka, and Sanjiv Kumar. Can looped transformers learn to implement multi-step gradient descent for in-context learning? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15130–15152. PMLR, 2024. URL https://proceedings.mlr.press/v235/gatmiry24b.html.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D. Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11398–11442. PMLR, 2023. URL https://proceedings.mlr.press/v202/giannou23a.html.

Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical Bayes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJ_UL-k0b.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cp5PvcI6w8_.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6. URL https://www.nature.com/articles/s41586-024-08328-6.

Giles Hooker. Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007. doi: 10.1198/106186007X237892.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998. URL https://www.cis.upenn.edu/~mkearns/papers/sq-journal.pdf.

Juno Kim and Taiji Suzuki. Transformers provably solve parity efficiently with chain of thought. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=E7HwPhfX1B.

Soo Min Kwon, Alec S. Xu, Can Yaras, Laura Balzano, and Qing Qu. Out-of-distribution generalization of in-context learning: A low-dimensional subspace perspective. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4285–4293, 2026.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Mary I. Letey, Jacob A. Zavatone-Veth, Yue M. Lu, and Cengiz Pehlevan. Pretrain–test task alignment governs generalization in in-context learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Florence Jessie MacWilliams and Neil J. A. Sloane. *The Theory of Error-Correcting Codes*. North-Holland, Amsterdam, 1977. ISBN 9780444851932.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Thomas Nagler. Statistical foundations of prior-data fitted networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25660–25676, 2023. URL https://proceedings.mlr.press/v202/nagler23a.html.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. doi: 10.1017/CBO9781139814782.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8298–8319. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.527. URL https://aclanthology.org/2023.findings-acl.527/.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-context learning through the Bayesian prism. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=HX5ujdsSon.

Jongho Park, Jaeseung Park, Zheyang Xiong, Nayoung Lee, Jaewoong Cho, Samet Oymak, Kangwook Lee, and Dimitris Papailiopoulos. Can mamba learn how to learn? a comparative study on in-context learning tasks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 39793–39812. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24j.html.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50817–50847, 2025. URL https://proceedings.mlr.press/v267/qu25d.html.

Zhenmei Shi, Junyi Wei, Zhuoyan Xu, and Yingyu Liang. Why larger language models do in-context learning differently? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 44991–45013. PMLR, 2024. URL https://proceedings.mlr.press/v235/shi24f.html.

Petar Veličković and Charles Blundell. Neural algorithmic reasoning. *Patterns*, 2(7):100273, 2021. doi: 10.1016/j.patter.2021.100273. URL https://arxiv.org/abs/2105.02761.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Kaiyue Wen, Huaqing Zhang, Hongzhou Lin, and Jingzhao Zhang. From sparse dependence to sparse attention: Unveiling how chain-of-thought enhances transformer sample efficiency. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fa6d4d2020aac4bd8f7cdb2771fc1ae2-Abstract-Conference.html.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

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

Table 2: Exact distributed-evidence support thresholds for squared risk at most 0.1, obtained from (13) and (14).

$d = 1/\rho^2$	n_{Bayes}	$n_{\mathcal{V}_1}$	$\Delta n_{0.1}$
2	5	9	4
3	9	18	9
4	12	27	15
6	19	45	26
8	26	63	37

A Experimental details and reproducibility

This appendix separates exact task-level calculations from the empirical records supporting the plots. Risk means squared error for sign-valued targets, so an uninformative zero predictor has risk one. Oracle risks use the same target convention as the learned predictor.

A.1 Task definitions and exact oracles

A.1.1 Factor controls and their oracle reference design

Let $Z_1, \ldots, Z_k$ be independent uniform signs. A complete-factor composition episode exposes the k factor outcomes and asks for $T = \prod_{j=1}^{k} Z_j$. The implementation samples rank-k ambient geometries conditioned on $k_{\min} = k$ with exactly k supports. Thus every support is required and there are no distractors; the embedding remains fresh across episodes. For $k \geq 2$, $R^\star = 0$, $R(\mathcal{V}_1) = 1$, and all target energy is in order k. This is the composition control used in Figure 3, with $k = 2, 3$.

A.1.2 Distributed evidence

For each episode, draw Z uniformly from $\{-1, +1\}$ and independent noise signs with $\Pr(\xi_i = 1) = (1 + \rho)/2$, then reveal $Y_i = Z\xi_i$ and predict the clean sign Z. All support and query inputs are the same zero scalar; only the observed evidence signs carry episode information. The shared-sign experiment varies $d = \rho^{-2} \in \{2, 3, 4, 6, 8\}$.

The posterior log odds are $\log \frac{\Pr(Z=1|Y)}{\Pr(Z=-1|Y)} = 2\,\mathrm{atanh}(\rho) \sum_i Y_i$, yielding (12). Let $a = \mathrm{atanh}(\rho)$ and $p = (1 + \rho)/2$. The exact Bayes risk is the finite binomial sum

$$R^\star(n, \rho) = \sum_{b=0}^{n} \binom{n}{b} p^b (1-p)^{n-b} \left[1 - \tanh^2\big(a(2b - n)\big)\right]. \tag{13}$$

The summand is even in $2b - n$, so averaging conditional on $Z = 1$ gives the same result as averaging over the symmetric mixture.

For the additive oracle, $\mathbb{E}Y_i = 0$, $\mathbb{E}Y_i^2 = 1$, $\mathbb{E}Y_iY_j = \rho^2$ for $i \neq j$, and $\mathbb{E}ZY_i = \rho$. The normal equations therefore give equal coefficients $\rho/[1 + (n-1)\rho^2]$ and the exact additive predictor and risk

$$P_1 m_{\mathrm{DE}} = \frac{\rho S_n}{1 + (n-1)\rho^2}, \qquad R_{\mathcal{V}_1} = \frac{1 - \rho^2}{1 + (n-1)\rho^2}. \tag{14}$$

For $n \geq 3$, the Bayes predictor is strictly outside $\mathcal{V}_1$: the outcome distribution has full support, and symmetry would force an additive representation to be proportional to S_n. But $\tanh(as)/s$ is strictly decreasing for positive s, and S_n has at least two distinct positive magnitudes. Hence $\Gamma_{>1} > 0$. The computation nonetheless needs only the shared statistic S_n.

The additive threshold at general risk ϵ is $\lceil (1 - \epsilon)(\rho^{-2} - 1)/\epsilon \rceil$. The learned points in Figure 3 are evaluated at these Bayes thresholds, namely $n = 5, 9, 12, 19, 26$. Each difficulty has a separately trained model; its support size is sampled uniformly from $\{0, \ldots, 64\}$ during training. Evaluation covers every integer in this range, with 2048 episodes per model seed and evaluation seed 91337.

Rates used in Table 1. For distributed evidence at fixed $\epsilon \in (0,1)$, let $d = \rho^{-2} \to \infty$. The additive formula gives $n_{\mathcal{V}_1} = \Theta(d)$. Majority vote has squared risk at most $4\exp(-n\rho^2/2)$ by Hoeffding's inequality, giving the Bayes upper bound $O(d)$. For the lower bound, let P_+ and P_- be the laws of the n observations conditional on $Z = +1$ and $Z = -1$. Their Kullback–Leibler divergence is $n\rho\log((1+\rho)/(1-\rho)) = O(n\rho^2)$. Moreover $R^\star = 1 - \mathbb{E}m_{\mathrm{DE}}^2 \geq 1 - \mathbb{E}|m_{\mathrm{DE}}| = 1 - \mathrm{TV}(P_+, P_-)$. Pinsker's inequality therefore requires $n = \Omega(d)$ to reach fixed risk below one. For latent codes, $R^\star \geq 1 - 2^{n-r}$ when $n \leq r$, since the support span contains at most 2^n of 2^r possible queries. Together with (10), this gives $n_{\mathrm{Bayes}} = \Theta(r)$. The complete-factor control has $n = k$ by design and no varying-support threshold; its additive risk remains one at that design point.

A.1.3 Latent-code episodes

The latent-code distribution is defined in Section 4. The stored natural-query banks use ambient dimension $D = 12$ and $(r, n) \in \{(2,3), (3,5), (4,6), (5,8), (6,9)\}$, where the final token is the query. Each rank and training-seed pair has 4096 episodes. Rank-six conditioned banks contain 1024 episodes per seed for each of $k_{\min} = 1$, $k_{\min} = 2$, $k_{\min} \geq 3$, and outside-span queries. The stored seeds are $17, 29, 43, 71, 101$, with separate episode-bank seeds for evaluation. Comparing models on a shared bank permits paired risk differences.

For each geometry, Gaussian elimination gives the support rank and a valid query route. Exact minimum order is obtained by enumerating support subsets (feasible at these support sizes) and taking the smallest subset yielding the query. All routes give the same clean label product. A zero query has known label $+1$; an outside-span query has Bayes mean zero. The additive oracle returns the known label for a zero or matching query and zero otherwise. Consequently, empirical oracle risk is the average of exact episode-level conditional risks, with no fitted oracle model.

Population Bayes risk can also be computed without sampling. If $p_t(s) = \Pr(\mathrm{rank}(U_{1:t}) = s)$, then $p_0(0) = 1$ and

$$p_{t+1}(s) = 2^{s-r}p_t(s) + (1 - 2^{s-1-r})p_t(s-1), \qquad R^\star(t) = 1 - \sum_{s=0}^{r} p_t(s)2^{s-r}, \tag{15}$$

with out-of-range probabilities zero. Together with $R_{\mathcal{V}_1}(t) = (1 - 2^{-r})^{t+1}$, this gives the exact integer thresholds plotted in Figure 2.

A.2 Interaction-spectrum computation with dependent outcomes

At a fixed geometry, enumerate distinct outcome states $y^{(s)}$ with their conditional probabilities $w_s > 0$. Let $W = \mathrm{diag}(w_s)$, $m_s = m_H(y^{(s)})$, and let Φ_q contain the evaluations of all monomials $\prod_{i \in S} Y_i$ with $|S| \leq q$. These monomials span all functions of at most q sign coordinates. If Q_q is the Euclidean orthogonal projector onto the column space of $W^{1/2}\Phi_q$, then

$$\Gamma_{>q}(H) = \|(I - Q_q)W^{1/2}m\|_2^2, \qquad I_q(H) = \|(Q_q - Q_{q-1})W^{1/2}m\|_2^2. \tag{16}$$

Use rank-revealing factorisation or a pseudoinverse when columns are dependent. Averaging over geometries gives the task spectrum. On latent codes, the character argument reduces this computation to $k_{\min}$. For distributed evidence, the state weights are

$$w(y) = 2^{-n-1}\left[\prod_i (1 + \rho y_i) + \prod_i (1 - \rho y_i)\right]. \tag{17}$$

Uniformly weighting the full Boolean cube would analyse a different problem.

A.3 Architectures and optimisation

Looped ICL uses an immutable labelled support memory, a query token, and six learned workspace tokens. Geometry is encoded by two linear maps with an intermediate GELU; label and token-type embeddings distinguish support, query, and workspace roles. Each shared processor round applies cross-attention to support memory, self-attention among the query and workspace tokens, and a GELU feedforward block. A direct query-to-support retrieval branch is fused with the final query

state before the scalar readout. The width is 64, each attention module has four heads, and the feedforward width is 160. The model uses pre-normalisation, zero dropout, residual scales initialised at 0.1, and a zero-initialised output head. The latent endpoint model has 89,953 trainable parameters. The main latent pipeline samples a loop count uniformly from $\{3, 4, 5, 6\}$ per training batch and evaluates at six loops. Factor controls, local-product controls, and distributed evidence use six loops throughout.

Set ICL is a query-aware Set Transformer with labelled support tokens and an unlabelled query token, full self-attention, width 64, four heads, GELU feedforward blocks, zero dropout, and a tanh readout. It uses four encoder layers and feedforward width 160. Its AdamW learning rate, batch size, and update budget match Looped ICL.

MAML and Meta-SGD use a scalar-output MLP with two hidden layers of width 64 and tanh activations. Both adapt all predictor parameters by support squared-loss gradients; outer differentiation includes the derivatives through these updates. MAML uses a fixed scalar inner learning rate. Meta-SGD learns one unconstrained step-size coefficient per predictor parameter, initialised at the selected inner rate. With $D = 12$, the predictor has 5057 parameters; Meta-SGD adds 5057 step sizes. The selected one-step models are reported in the appendix. Repeating their update fields at evaluation is a separate diagnostic (Appendix B).

Outer optimisation uses AdamW with its implementation defaults for moments and numerical stabilisation, a global gradient-norm clip of one, and at most 20,000 updates. Looped ICL uses learning rate 3×10^{-4}, weight decay 10^{-4}, and 96 episodes per update. Main latent and distributed-evidence adapters use batches of 32 and zero weight decay; their outer rate is selected as described below. The historical factor, route-revealed, and local-control runner specifies weight decay 10^{-4} also for adapters, while preserving the selected adapter outer rate. All losses are mean squared query error unless an explicitly stated auxiliary term is added.

A.4 Training distributions and model selection

For latent-code endpoint training, ranks are uniform on $\{2, 3, 4, 5, 6\}$ and support sizes are uniform on $\{0, \ldots, 9\}$. Fresh tasks and full-rank embeddings are sampled for every episode; binary geometry is encoded as signed bits. With probability 0.35, the query is natural and uniform. Otherwise, the sampler chooses uniformly among the nonempty strata of minimum orders $1, \ldots, r$ and the outside-span stratum, then uniformly within the chosen stratum. This curriculum supplies examples of different evidence requirements without revealing their route to the endpoint model. Evaluation uses natural queries unless a minimum-order condition is stated.

Validation occurs every 1000 updates on a fixed bank with 512 episodes per rank, ranks 2–6, $n = \lceil 1.5r \rceil$, and base seed 51337. The mean validation risk selects the checkpoint. Early stopping uses minimum step 8000, patience five validation checks, and minimum improvement zero. The auxiliary-control runner uses validation base seed 61337, 512 episodes for each factor control and per rank for route-revealed validation, and the same validation interval and stopping settings. Local-product validation repeats all four possible sign pairs equally to reach 512 examples. Distributed evidence uses 1024 validation episodes with support sizes 0–64, base seed 51337, and the same 1000-step interval and stopping rule. Evaluation batches have 128 episodes.

For the main endpoint sweep, ranks 2–6 participate in both final training and checkpoint validation. The rank sweep therefore measures increasing task difficulty within the training range. The rank restrictions used in hyperparameter search below do not turn it into an unseen-rank generalisation experiment.

A.5 Representation ladder

All conditions use $D = 12$, ranks sampled uniformly from 2–6, support counts sampled uniformly from 0–9, and the endpoint query curriculum above. Canonical inputs use $[I_r \ 0]$. Shared inputs use the first r rows of one random invertible $D \times D$ matrix per training seed, held fixed throughout training and validation; the ranks therefore share nested row spaces. Fresh inputs resample the embedding each episode. Shared-to-unseen evaluation reuses the shared-trained model on four unseen fixed embeddings per rank and seed; it performs no additional training or model selection. The ladder inherits the endpoint optimisation and validation settings. Set ICL provides the second

contextual-architecture comparison. Natural evaluation uses $n = \lceil 1.5r \rceil$, 4096 episodes per rank and seed, with the same latent episodes rendered under each representation. Oracle risks are the exact population quantities. For unseen-basis evaluation, metrics are averaged over test bases within a training seed before computing five-seed means and Student-t intervals. Bases are not treated as independent training replicates.

A.6 Rank-general recurrent transition experiment

The recurrent experiment uses ambient dimension $D = 12$, five training seeds, and a mutable-state model distinct from the endpoint contextual learners.

State, architecture, and supervision. The mutable state comprises up to 16 padded support rows and label bits, locked-row flags, a query residual, an accumulated label bit, completed-coordinate flags, and a halt bit. Padding and normalised row-order keys are immutable; keys break teacher pivot ties. Geometry, binary values, row keys, and token roles are embedded with linear/GELU maps. The processor is a tied four-layer bidirectional Transformer with six workspace tokens, four heads, width $h \in \{64, 128, 256\}$, feedforward width $2.5h$, pre-normalisation, and zero dropout. Linear heads predict mutable fields and pivot row/coordinate logits. State updates are thresholded at 0.5, with straight-through gradients during training, and re-encoded each round. A two-hidden-layer GELU/tanh readout predicts the query sign from the final state. Test execution contains no oracle update, pivot, span test, or route. A transition contains four layers; its count is not directly comparable to Looped-ICL rounds.

The teacher performs binary elimination: choose the first eligible unfinished coordinate and unlocked row, XOR the pivot into eligible rows and the query, update label bits and control flags, and halt when no pivot remains. The process-supervised objective is

$$\mathcal{L} = \mathcal{L}_{\text{answer}} + \mathcal{L}_{\text{state}} + 0.25\mathcal{L}_{\text{pivot}}.$$

Answer loss is query MSE. State loss averages binary cross-entropy across fields and active transitions, excluding padding; changing bits have weight three and unchanged bits weight one. State targets follow the canonical teacher trajectory at the same transition index, including during predicted rollouts. Pivot loss averages row and coordinate cross-entropies when a valid teacher pivot exists. Auxiliary loss weights remain fixed for 30,000 updates. Teacher forcing substitutes the teacher state after an active transition with probability $\max(0, 1 - t/15000)$ at update t; it is absent in the second half of training and at evaluation. The width-64 supervision ablations remove teacher forcing, pivot loss, or state loss as indicated in the figure.

Training distribution, controls, and optimisation. The witness and plain endpoint-only control (variant C) sample ranks uniformly from $\{3, 4, 5\}$ and support sizes uniformly from $\{4, \ldots, 16\}$, with fresh full-rank embeddings and tasks. Queries are natural with probability 0.35; otherwise the sampler selects among feasible minimum-order strata and the outside-span stratum as in Appendix A.4. For each episode, training recurrence is uniform on $\{r, \ldots, 8\}$. At each width, variant C has the same state representation, architecture, and readout as A, but its only loss is final query MSE and it never uses teacher forcing.

The stronger endpoint-only control (C3) adds gated residual state updates, with gate weights initially zero and biases -2, and averages answer MSE at $T \in \{r, r + 2, r + 4\}$. It adds 0.05 times the squared difference between predictions at depths $r + 4$ and $r + 5$. It samples ranks $\{2, 3\}$ through update 8,000, $\{2, 3, 4\}$ through 18,000, then $\{3, 4, 5\}$ with 10% rank-two replay through 30,000. It receives no oracle states, pivots, or routes. The C3 comparison therefore includes architectural and curriculum changes; A versus C is the matched architecture comparison. Figure 5a uses A versus C3 at width 64; panel (b) uses A versus C at width 64; panel (c) uses A versus C at width 256.

All three conditions use 30,000 AdamW updates, batches of 96 episodes, learning rate 3×10^{-4}, weight decay 10^{-4}, default AdamW moments and stabilisation, and global gradient-norm clipping at one. Early stopping is disabled. Training seeds are $17, 29, 43, 71, 101$.

Validation, held-out ranks, and test banks. Checkpoint selection minimises mean validation MSE on ranks $3, 4, 5$ only, using 512 fixed natural episodes per rank and checks every 1,000 updates. Validation uses $n = \lceil 1.5r \rceil$ and $T = r$. Ranks 6–9 are excluded from both training and checkpoint

Table 3: Adapter search grid. For Meta-SGD, the inner rate is the initialisation of its learned coordinate-wise step sizes.

ID	Steps	Inner rate	Outer rate	ID	Steps	Inner rate	Outer rate
c01	1	0.01	3×10^{-4}	c06	4	0.10	3×10^{-4}
c02	1	0.03	3×10^{-4}	c07	1	0.03	10^{-3}
c03	1	0.10	3×10^{-4}	c08	4	0.03	10^{-3}
c04	4	0.01	3×10^{-4}	c09	8	0.01	3×10^{-4}
c05	4	0.03	3×10^{-4}	c10	8	0.03	3×10^{-4}

selection. This is held-out intrinsic-rank transfer at fixed ambient dimension; the test support lengths remain within the training range.

The paper-facing test bank contains 4,096 natural episodes per rank per training seed, for ranks 3–9, using $n = \lceil 1.5r \rceil$ and independent evaluation episodes. Models and oracles share each bank. The primary rule is $T = r$, not a depth chosen using held-out performance. The depth sweep retains $T \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 16\}$ with the same selected checkpoint. Minimum-order diagnostics partition natural episodes by in-span $k_{\min} = 1, 2, 3, 4+$. Seed-level ratios and 95% Student-t intervals follow Appendix A.9. Supplementary full-grid diagnostics use 256 episodes per cell.

A.7 Adapter hyperparameter search

The main adapter search evaluates the ten recipes in Table 3 separately for MAML and Meta-SGD. Each recipe trains on order-two factors, order-three factors, and rank-four latent codes, with two tuning seeds (7301 and 9103): 60 runs per method. Each run allows 5000 updates with batch size 32, validation every 500 steps on 512 episodes (base seed 81337), and early stopping after at least 1500 steps with patience four and minimum improvement zero. Selection first averages the best validation risk over the two seeds for each task, then gives the three tasks equal weight. Candidate identifier breaks ties deterministically. The final runner requires the selected configuration and freezes one recipe per method across the latent rank sweep and the factor/route controls. Final test banks are excluded from this selection.

The distributed-evidence search uses the same ten recipes and two tuning seeds, with difficulties $d \in \{2, 4, 8\}$ replacing the three main tasks. It has the same 5000-update, 500-step validation, and early-stopping budgets; validation has 512 episodes spanning support sizes 0–64. The selector averages over tuning seeds within each difficulty and then equally over the three difficulties. Its output is a selected-configuration file. The final distributed-evidence runner reads its paper configuration directly; that configuration specifies one inner step, inner rate/initialisation 0.1, and outer rate 3×10^{-4} for both adapters. Looped ICL uses the fixed architecture and optimiser recipe above. Thus adapter search includes longer inner horizons, but the repeated-update diagnostic does not represent independently trained multi-step checkpoints.

A.8 Routing and structural interventions

The hidden, random, and revealed route-cue conditions use the same $D + 2$ Looped-ICL input interface, initialisation, optimiser, episode distribution, geometry-only in-span flag, and endpoint loss. Hidden sets the route channel to zero; random supplies geometry-keyed Bernoulli cues independent of the query route; revealed supplies a valid minimum-cardinality route computed from geometry. Labels never enter cue construction. Training uses ranks 2–6; evaluation uses rank six, $n = 9$, and natural episodes partitioned by $k_{\min} = 1, 2, 3+$. Thus the random condition controls for merely adding an extra input channel.

Two auxiliary structural conditions use the original D-dimensional interface. Graph-SVOR adds local consistency probes

$$f_D(a \oplus X_j) \approx f_D(a)Y_j, \qquad f_D(0) \approx 1,$$

where the anchor a is generated from observed support geometry; probe construction does not solve the endpoint route. Span residual instead trains a geometry-only route head with probabilities p_i on

auxiliary in-span queries. Its soft XOR prediction for coordinate j is

$$\pi_j(p; X) = \frac{1 - \prod_{i:X_{ij}=1}(1 - 2p_i)}{2}.$$

The auxiliary binary-cross-entropy loss acts through the shared geometry encoder; in the reported condition, predicted route scores are not fed directly to the endpoint predictor.

For the local-composition positive control, training samples independent sign pairs (z, y) with target zy. At evaluation, an exact route is supplied and the learned multiplication rule is repeatedly applied along its support labels. This tests learned composition given routing, not autonomous route construction or elimination.

A.9 Seed aggregation and confidence intervals

All main learner summaries use seeds $17, 29, 43, 71, 101$. Shared natural episode banks support paired model/oracle comparisons. Each seed has 4096 natural episodes per rank. Conditioned rank-six banks have 1024 episodes per stratum. Figure 14 partitions the natural banks by minimum order, with seed-level episode counts retained in the data tables. Means give equal weight to training seeds, rather than weighting seeds by their realised stratum sizes.

For each seed, compute risk and paired oracle risks, then form

$$\eta_s = \frac{R_{\mathcal{V}_1,s} - R_{\mathrm{method},s}}{R_{\mathcal{V}_1,s} - R^\star_s}. \tag{18}$$

Average these ratios across seeds. Distributed evidence uses its exact population oracle risks with the seed-level learned risks. The two-sided 95% Student-t interval for a seed-level statistic u_s is

$$\bar{u} \pm t_{0.975,4}\frac{s_u}{\sqrt{5}}, \qquad s_u^2 = \frac{1}{4}\sum_{s=1}^{5}(u_s - \bar{u})^2, \quad t_{0.975,4} \simeq 2.776. \tag{19}$$

These intervals quantify variation across the five training-seed summaries; episodes and repeated inner steps are not independent training replicates. Intervals are pointwise and are not clipped to $[0, 1]$.

A.10 Gradient-access quantities and estimation

For a conditional target v_H, the geometry-specific numerator is

$$g_H(v; \theta) = \sum_t w_t v_H(y^{(t)}) \nabla_\theta F_\theta(H, y^{(t)}), \qquad A_H(v; \theta) = \|g_H(v; \theta)\|_2^2. \tag{20}$$

Compute the conditional average vector before squaring its Euclidean norm, then average over geometries. Latent-code access enumerates all 2^r latent tasks for each geometry, inducing exactly the conditional label law even when several tasks produce identical support outcomes. No noisy Monte Carlo square is used for this conditional average.

Figure 15a uses 256 fixed geometries per rank and minimum-order stratum (base seed 271828), with the same bank across five model initialisations and six evaluation loops. It reports $\overline{A_H(m; \theta)}/\overline{\|m_H\|_2^2}$, a ratio of geometry means. The analytic control has eight heads, hidden width 64, GELU activation, value gain $\gamma = 0.25$, analytic radius one, and model seed offset 314159. Its local reference is eight times the squared four-sign correlation in (61), one contribution per readout head, giving 1.2179×10^{-3}. The plotted order-two slice has target energy one. The access–competence panel pairs these initialisation summaries with natural-bank order-two gains using the same training seeds.

For training trajectories, use rank six, nine supports, minimum order two, 128 shared geometries (base seed 223607), and fixed checkpoints at steps $0, 2000, 4000, 6000, 8000$. These are milestone checkpoints, independent of best-validation checkpoint selection. Set $v_H = m_H - F_\theta(H, \cdot)$ and hold this residual fixed when differentiating the prediction. To make the normalisation explicit, Figure 15c reports

$$\mathfrak{a}_{\mathrm{res}}(\theta) = \frac{\overline{A_H(m_H - F_\theta; \theta)}}{\overline{\|m_H - F_\theta\|_2^2}}, \tag{21}$$

computed from the seed-level residual numerator and energy. No Jacobian-energy factor is included. The plot shows the mean of the five seed-level ratios and each individual trajectory. Some Student-t intervals cross zero and cannot be displayed fully on a logarithmic axis. Fixed geometry banks make this uncertainty conditional on the bank. Access remains sensitive to parameter coordinates and feature scaling; it is not a global difficulty score or a convergence rate.

B Supplementary learning diagnostics

B.1 Broader learner comparison and repeated adapter updates

Figure 6 includes MAML and Meta-SGD. Their weak order-three composition performance prevents attributing their latent-code failure specifically to route discovery. Repeating the learned update fields can reduce support loss without improving held-out higher-order prediction (Figure 7); these are evaluation-time update sweeps, not independently trained multi-step methods.

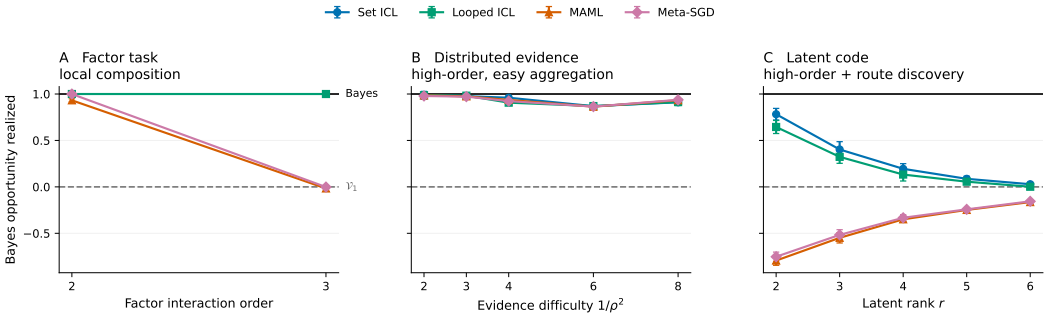


Figure 6: Contextual and parameter-adaptation learners across the three task families. Means and pointwise 95% intervals use five training seeds.

The repeated updates confirm that the adapters are not simply inactive. On the order-three factor task, eight updates reduce support loss to about 0.071 for MAML and 0.16 for Meta-SGD, while held-out query loss remains near or above one. On rank-six latent episodes with $k_{\min} \geq 3$, support loss falls from 0.999 to 0.092 for MAML and from 0.994 to 0.071 for Meta-SGD, while query losses reach 1.42 and 1.44. Thus within-task optimisation can fit the observed supports without recovering the episode-level rule needed at a fresh query. This is an empirical limitation of these learned update rules, not a claim that nonlinear gradient adaptation is restricted to $\mathcal{V}_1$.

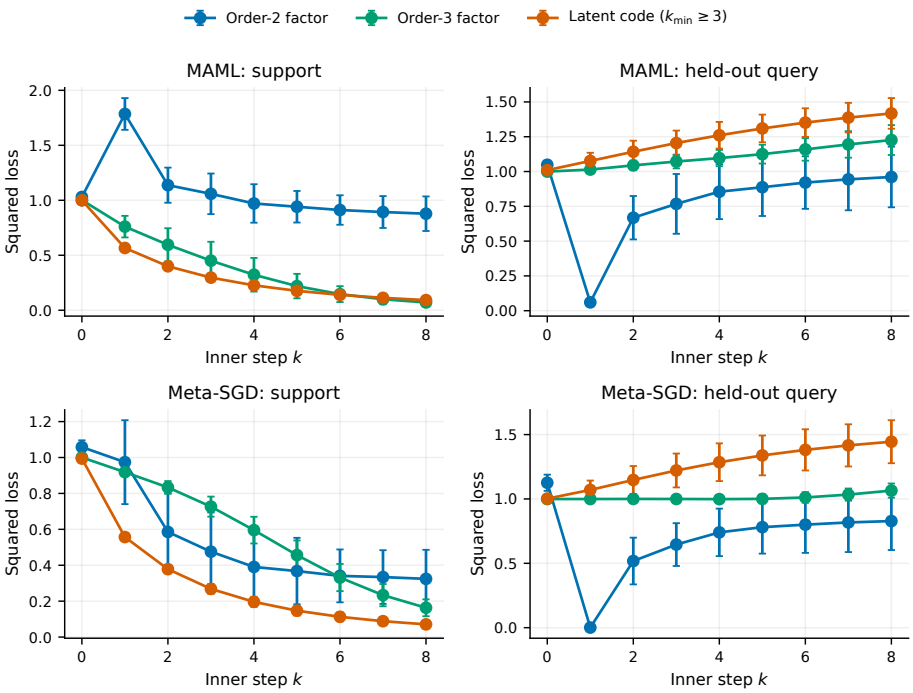


Figure 7: Support fitting and query risk as selected MAML and Meta-SGD update fields are repeated at evaluation. Fitting the support outcomes does not ensure successful held-out composition.

B.2 Process supervision, width, and execution

The width sweep retains endpoint failure at widths 64, 128, and 256 while process-supervised models transfer to larger ranks (Figure 8). The width-64 supervision ablation removes teacher forcing, state supervision, or pivot supervision. Full supervision is strongest; state-only and no-teacher-forcing conditions retain smaller gains. Pivot-only supervision does not reproduce successful execution. These results concern acquisition under the specified budgets, not a necessity theorem for any one auxiliary objective.

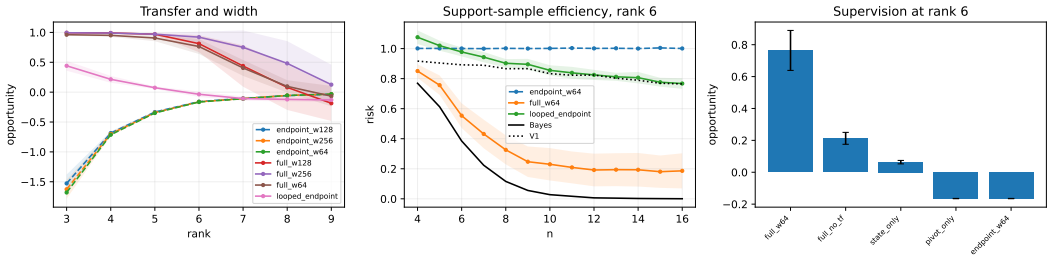


Figure 8: Original process-supervision diagnostics: transfer across widths, support-size risk at rank six, and supervision ablations. The middle panel also supplies Figure 5b.

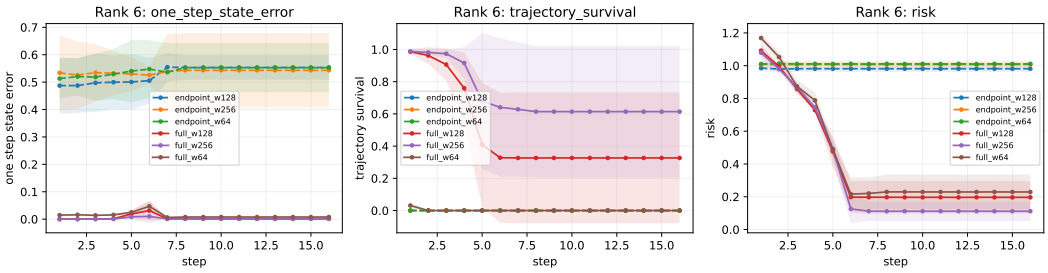


Figure 9: Learned execution at rank six. One-step error, trajectory survival, and final risk need not coincide: useful predictions can survive imperfect agreement with the teacher's full state sequence.

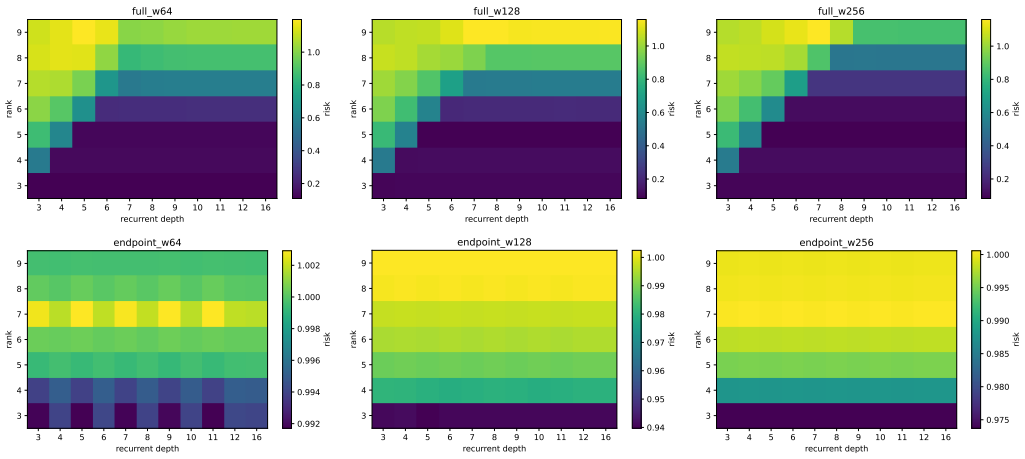


Figure 10: Depth sweeps across widths: process-supervised models (top) and matched endpoint models (bottom). These original diagnostic panels use *different colour scales*; the shared-scale comparison is Figure 5c.

Retention after removing the scaffold. Continuation from an acquired process-supervised model tests retention, not acquisition from an endpoint-only initialisation. Across the tested zero/fast/medium/slow removal schedules, higher-order recovery deteriorates; continued full supervision retains it (Figure 11). The endpoints preserve much order-one retrieval but lose composition of multiple support labels. Figure 12 tracks the associated decline in pivot and state prediction and final recovered opportunity. This makes retention a substantive failure mode even though the acquired model already executes without a teacher at test time.

These trajectories do not show that the endpoint population gradient points away from a correct algorithm. Continued training changes parameters under finite-batch optimisation, an imperfect learned transition, and potentially regularisation and optimiser-state effects. The architecture uses zero dropout and the reported recurrent training uses AdamW weight decay 10^{-4} (Appendix A.6); the continuation plots do not isolate the contribution of weight decay. No causal attribution to regularisation, gradient direction, or a particular failure sequence is therefore warranted. The supported conclusion is loss of acquired higher-order behaviour under the tested continuation protocols.

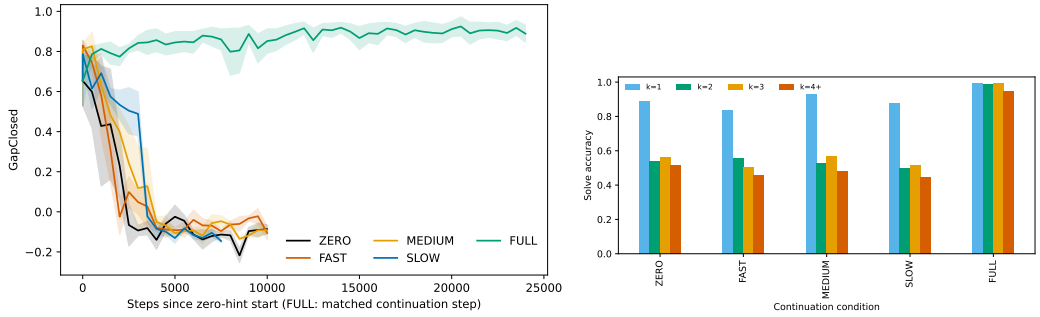


Figure 11: Scaffold removal: opportunity after continuation (left) and final in-span accuracy by minimum route order (right). Continued endpoint training can erode an autonomously executable computation.

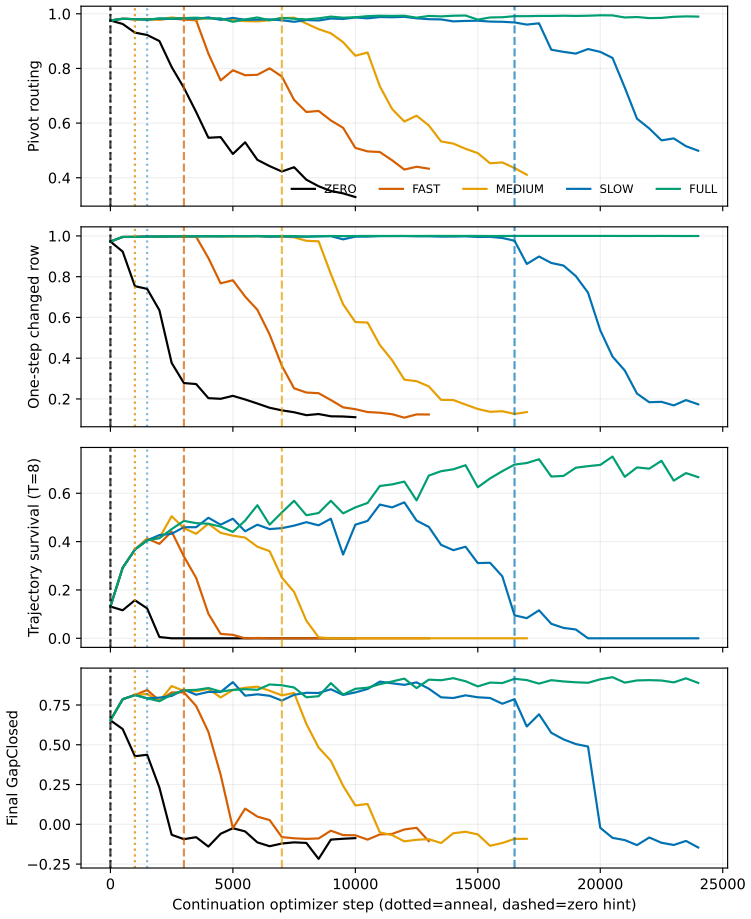


Figure 12: **Retention during continued training.** From top to bottom: pivot-routing accuracy, one-step changed-row accuracy, trajectory survival at $T = 8$, and final recovered opportunity. Dotted vertical lines mark annealing transitions and dashed lines mark zero auxiliary hint weight. Curves show deterioration under the tested removal schedules, without identifying its causal mechanism.

B.3 Routing and structural interventions

Conditioning on $k_{\min}$ separates the number of labels that must be composed from the remaining episode geometry. Figure 13 shows that endpoint Looped ICL loses higher-order competence as rank grows even at fixed composition order: for $k_{\min} = 2$, gain over $\mathcal{V}_1$ falls from 0.74 at $r = 2$ to 0.02 at $r = 6$; for $k_{\min} \geq 3$, it falls from 0.35 at $r = 3$ to 0.01 at $r = 6$. Thus composition order alone does not explain the latent-code failure.

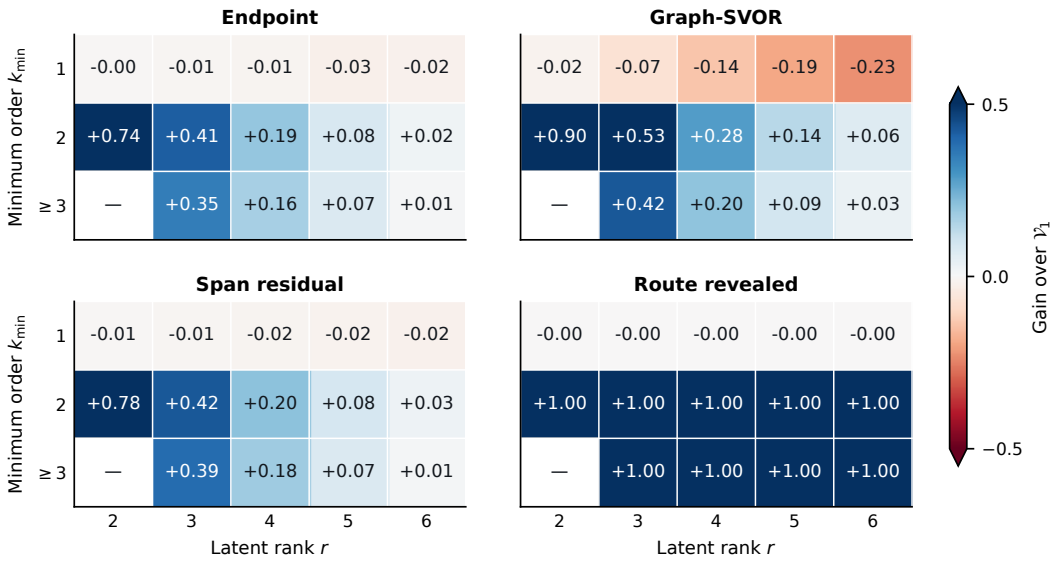


Figure 13: **Structural interventions on latent codes.** Cells report conditional gain $R_{\mathcal{V}_1} - R(F_\phi)$ by rank and minimum explanation order. Graph-SVOR and span-residual supervision give partial improvements, whereas supplying an exact valid route restores almost all higher-order gain. Near-zero gain is optimal at order one because $\mathcal{V}_1$ is already Bayes-optimal there.

Graph-SVOR and span residual provide weaker structural signals without revealing the endpoint route. They shift the failure boundary only modestly. Exact route revelation is qualitatively different: it identifies which support labels should be composed while leaving the endpoint target unchanged. At rank six, realised Bayes–$\mathcal{V}_1$ opportunity increases from 0.002 under ordinary endpoint training to 0.998 with the route supplied. This shows that the contextual learner can perform the required composition when the episode-specific organisation of the evidence is made explicit.

The matched extra-channel control rules out the simpler explanation that route revelation merely gives the model additional inputs. Hidden, random, and revealed conditions use the same interface; only the valid route restores higher-order performance.

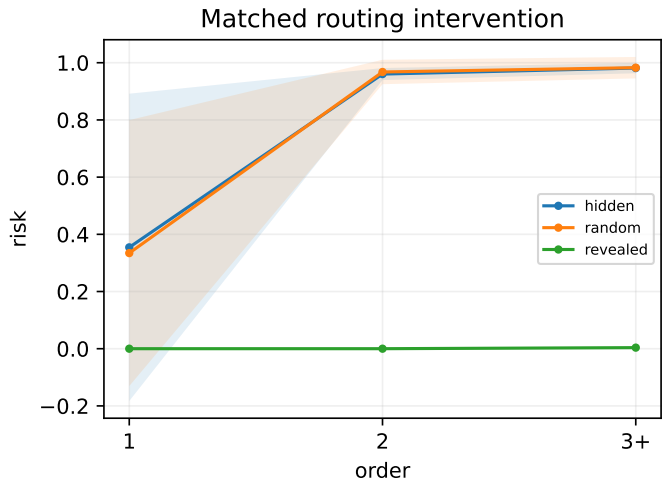


Figure 14: **Matched route-information control.** Rank-six risk by minimum route order under hidden, random, and revealed cues, using the same input interface.

B.4 Exploratory gradient diagnostics

These diagnostics are supplementary observations, not premises of the representation or sample-efficiency results. The analytic model below is distinct from the trained contextual architectures. Fixed-order target access decreases with rank and is descriptively associated with competence in these runs; rank and training condition also vary, so this is not a controlled causal comparison or an independently validated predictor of learnability. Revealing routes changes residual access during training. These parameterisation-dependent diagnostics do not verify the restricted theorem's assumptions for the trained contextual architectures or establish a convergence rate.

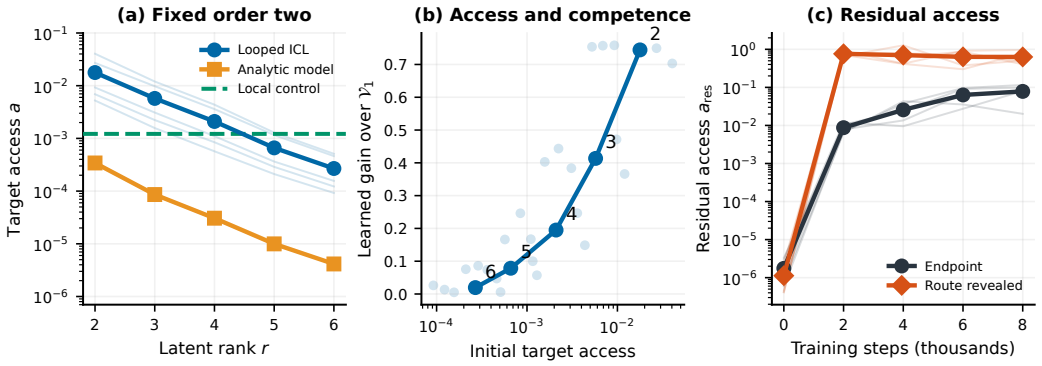


Figure 15: Supporting gradient diagnostics: fixed-order access, its association with learned competence, and endpoint versus route-revealed residual access. Axes and definitions follow Appendix A.10.

The analytic control provides the complementary order sweep predicted by the restricted theorem. As exact $k_{\min}$ increases, endpoint target access falls sharply, while the exposed local multiplication objective retains constant access (Figure 16). This varies the interaction order itself, whereas Figure 15a fixes $k_{\min} = 2$ and varies rank. The two diagnostics therefore probe distinct sources of weak endpoint alignment. Neither defines a parameterisation-invariant difficulty measure.

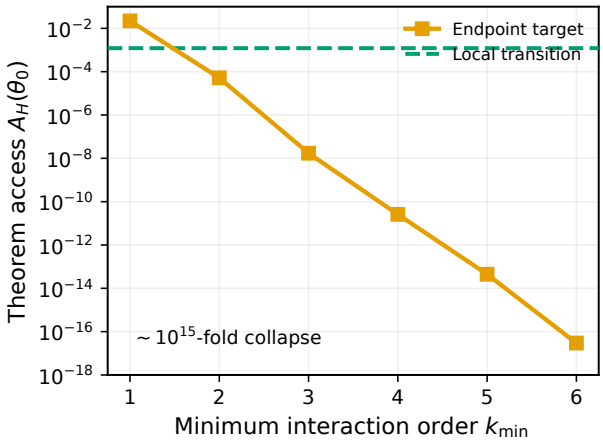


Figure 16: **Analytic access versus interaction order.** Endpoint target access declines steeply with exact $k_{\min}$ in the restricted analytic model, while access to the exposed local multiplication target remains constant.

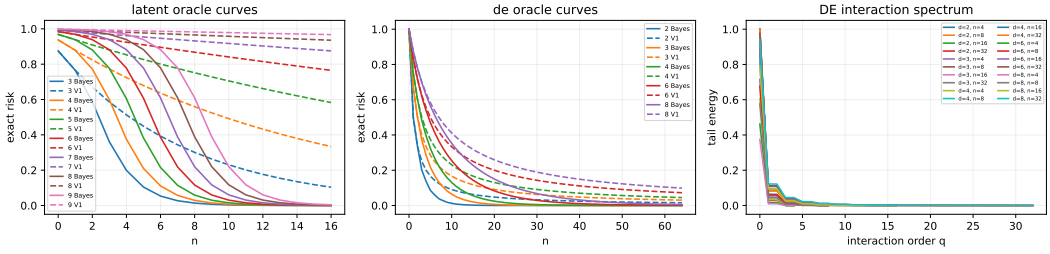


Figure 17: Exact oracle curves and distributed-evidence interaction spectra. The shared aggregate can realise a nonadditive Bayes map.

C PROOFS FOR THE HIERARCHY AND FIXED-GEOMETRY ADAPTATION

C.1 PROJECTION GEOMETRY AND DISTRIBUTION DEPENDENCE

For any square-integrable predictor measurable in (H, Y), conditional expectation gives $\mathbb{E}[(Y_{\mathrm{q}} - m_\Pi)(m_\Pi - g)] = 0$. Expanding the squared error yields

$$R(g) = R^\star + \|g - m_\Pi\|_2^2. \tag{22}$$

The definition of $\mathcal{V}_q$ includes closure, so the Hilbert projection theorem provides a unique $P_q m_\Pi$ as an L^2 equivalence class. It minimises the second term, proving Theorem 2.

Nestedness implies $P_q P_{q-1} = P_{q-1}$ and makes $P_q - P_{q-1}$ an orthogonal projector. Since $\mathcal{V}_n$ contains every function of (H, Y), its increments decompose m_Π, proving (3). Dependence between labels may make individual subset representations nonunique. The order projections remain unique under the actual episode law. In particular, low order is a property of the function on the distribution's support, not the degree of an arbitrary extension to unobserved outcome combinations.

Every function in $\mathcal{V}_q$ may use arbitrary measurable geometry coefficients. For the finite conditional state spaces used here, the projection can therefore be computed separately at each H and integrated, as in (16). For general outcomes, the global closed-space definition suffices for the risk identity and requires no subset-wise orthogonalisation. This distinguishes it from stronger dependent-variable ANOVA constructions (Hooker, 2007; Chastaing et al., 2012).

Proposition 7 (An approximate interaction restriction)**.** *Let $d_q(g) = \mathrm{dist}(g, \mathcal{V}_q)$. Then*

$$R(g) - R^\star \geq (\sqrt{\Gamma_{>q}} - d_q(g))_+^2. \tag{23}$$

If $\sup_{g \in \mathcal{A}} d_q(g) \leq \epsilon_{\mathcal{A}}$, the same bound holds for $\inf_{g \in \mathcal{A}} R(g) - R^\star$ with $d_q(g)$ replaced by $\epsilon_{\mathcal{A}}$.

Proof. Distance to a closed set is 1-Lipschitz, hence $\sqrt{\Gamma_{>q}} \leq \|m_\Pi - g\|_2 + d_q(g)$. Apply (22) and take an infimum for the class bound. □

Large distance from $\mathcal{V}_q$ alone does not imply useful higher-order prediction: the additional component must align with the Bayes target.

C.2 FINITE-STEP FIXED-GEOMETRY ADAPTATION

Rewrite (5) as

$$w_{t+1} = (I - M_{t,H} A_H) w_t + M_{t,H}(B_H Y + c_H). \tag{24}$$

Set $\alpha_0 = w_0$ and $C_0 = 0$. Inductively,

$$\alpha_{t+1} = (I - M_{t,H} A_H)\alpha_t + M_{t,H} c_H, \tag{25}$$

$$C_{t+1} = (I - M_{t,H} A_H) C_t + M_{t,H} B_H. \tag{26}$$

All coefficients remain label-independent, so $w_K = \alpha_K + C_K Y$. The affine query head preserves this form. For a square-integrable resulting predictor, Theorem 2 gives the excess-risk bound, proving Theorem 3. Any L^2 limit of such predictors also belongs to $\mathcal{V}_1$, because the space is closed.

Assumptions that matter. The number of updates and their preconditioners may depend on H, but not on observed labels or label-dependent iterates. A label-dependent stopping rule or state-dependent preconditioner can invalidate the affine recurrence. Nonlinear support-feature adaptation and a nonlinear query head can also create higher-order dependence. Learning a representation across tasks is compatible with this theorem when that representation stays fixed during within-task quadratic adaptation; it does not make the theorem a statement about all of MAML or Meta-SGD.

A Gaussian positive example. Let $\theta \sim \mathcal{N}(\mu_0, \tau^2)$ and $Y_i, Y_{\mathrm{q}} \mid \theta \overset{\mathrm{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$. Conjugacy gives

$$m(Y) = \frac{\tau^{-2}\mu_0 + \sigma^{-2}\sum_i Y_i}{\tau^{-2} + n\sigma^{-2}} \in \mathcal{V}_1. \tag{27}$$

One gradient step on $\frac{1}{2n}\sum_i (w - Y_i)^2$, starting at μ_0 with step size $n\sigma^{-2}/(\tau^{-2} + n\sigma^{-2})$, returns this exact Bayes predictor. This is an elementary conjugate-inference example (Diaconis & Ylvisaker, 1979) with no higher-order opportunity to miss.

C.3 How a nonlinear query head creates interactions

For one step on the half mean squared support loss of a differentiable model,

$$\theta^+(Y) = \bar{\theta}(H) + \sum_i Y_i v_i(H), \qquad v_i(H) = \frac{\eta}{n} \nabla_\theta f_{\theta_0}(X_i). \tag{28}$$

The final predictor $g(Y) = f_{\theta^+(Y)}(X_{\mathrm{q}})$ need not be additive. If the query map is q times differentiable on this update region, then for distinct $i_1, \ldots, i_q$ the chain rule gives

$$\frac{\partial^q g}{\partial Y_{i_1} \cdots \partial Y_{i_q}} = D_\theta^q f_{\theta^+(Y)}(X_{\mathrm{q}})[v_{i_1}, \ldots, v_{i_q}]. \tag{29}$$

Its magnitude is at most $M_q \prod_j \|v_{i_j}\|_2$ when the parameter derivative has operator norm at most M_q. This local differential identity explains an escape from additivity. It is not a bound on the full interaction tail for arbitrary nonsmooth networks or a guarantee that the generated interactions are useful.

D Latent-code finite-sample theory

D.1 Representation invariance and coordinate ambiguity

Proof of Proposition 4. Full row rank makes $u \mapsto uG$ injective. Hence any XOR of support inputs is zero (or equals the query) before embedding if and only if the same relation holds afterwards. In particular the route sets and their minimum cardinalities agree, including the zero-query and outside-span cases.

For fixed latent geometry, the binary vector of support labels is the image of a uniform task code under a linear map. It is uniform on its image; this image is characterised exactly by the linear dependencies among the support rows. Those dependencies are preserved by embedding. The same holds after including the query row. Thus the conditional joint law of support and query labels is determined by the observed relations and equals its latent-geometry counterpart. Conditioning on an observed geometry while integrating out an unknown embedding does not alter that law: every compatible latent geometry has the same relations and therefore the same conditional law.

At fixed geometry, the subspace spanned by functions of at most q support outcomes depends only on this conditional law. Its Bayes target, inner product, and orthogonal projections are identical in the two representations. All conditional projection energies and risks therefore agree. Averaging over coupled uniform latent episodes gives the population statements, regardless of whether an independent full-rank embedding is fixed or freshly sampled. Equality at every support size gives equality of fixed-risk thresholds. The assumptions on the task prior and independence matter: a representation that encodes additional task information is outside the proposition. □

Proof of Proposition 5. A full-row-rank G_0 has a right inverse over $\mathbb{F}_2$, proving shared canonicalisation. If a single R decoded all latent inputs for both G and AG, it would satisfy $GR = I_r$ and $AGR = I_r$; substituting gives $A = I_r$, a contradiction. This excludes a universal linear inverse for a nontrivial basis-randomised family, not an input-dependent algorithm. For the change of basis in (8), $(U_i A)(A^{-1}G) = U_i G$ and $(U_i A)(zA^{-T})^T = U_i z^T$. Uniform latent inputs and task codes are invariant under this invertible change. The observed episode therefore has indistinguishable descriptions with different latent coordinates. Its relation class, and the label product along any valid route, are nevertheless unchanged. □

D.2 Conditional characters and the exact spectrum

Fix a label-free geometry and choose latent coordinates consistent with its linear relations. Define the syndrome map and the conditional label group by

$$T_H : \mathbb{F}_2^n \to \mathbb{F}_2^r, \quad T_H c = \bigoplus_{i : c_i = 1} U_i, \qquad C_H = \{((-1)^{\langle z, U_i \rangle})_{i=1}^n : z \in \mathbb{F}_2^r\}. \tag{30}$$

The support labels are uniform on C_H: every attainable label vector has the same number of task-code preimages. For $c \in \mathbb{F}_2^n$, let $\chi_c(Y) = \prod_{i:c_i=1} Y_i$. Then

$$\mathbb{E}[\chi_c \chi_{c'} \mid H] = \mathbf{1}\{T_H c = T_H c'\}. \tag{31}$$

Thus different characters are indexed by syndromes, with repeated subset representations corresponding to the same character. Their minimum degree is

$$q_H(c) = \min\{\mathrm{wt}(a) : T_H a = T_H c\}. \tag{32}$$

Changing the latent basis preserves C_H and all these relations. The proof therefore also applies when the random embedding G is unobserved.

If U_{q} is in the support span, choose c with $T_H c = U_{\mathrm{q}}$. The query is determined by the observed labels as $Y_{\mathrm{q}} = \chi_c(Y)$, independently of which valid c is chosen. Otherwise there exists a task-code displacement orthogonal to every support but not to the query. Pairing tasks differing by that displacement leaves all observed labels unchanged and flips the query. Its conditional mean is then zero. This proves the Bayes predictor in Theorem 6.

Any function of Y_S on the sign cube admits a multilinear expansion in χ_a supported within S; restriction to C_H preserves that expansion. If $k_{\min} > q$, none of these characters equals the query character when $|S| \le q$, so the query is orthogonal to every such function. If $k_{\min} \le q$, a valid subset product represents it exactly. Geometry coefficients are unrestricted, hence

$$(P_q m)(H, Y) = \mathbf{1}\{k_{\min}(H) \le q\}\, m(H, Y). \tag{33}$$

The target has conditional squared norm one on in-span geometries and zero outside. Squaring successive projection increments and averaging proves $I_q = \Pr(k_{\min} = q)$, including $I_0 = \Pr(U_{\mathrm{q}} = 0) = 2^{-r}$. The conditional risk is one precisely when the query cannot be represented at the allowed order. This proves all identities in (9), including $R^\star = \Pr(k_{\min} = \infty)$.

D.3 Support-sample complexity

If n supports fail to span $\mathbb{F}_2^r$, some nonzero vector is orthogonal to all of them. For each such vector this occurs with probability 2^{-n}. A union bound gives

$$R^\star \le \Pr(\mathrm{rank}(U_{1:n}) < r) \le (2^r - 1)2^{-n} < 2^{r-n}. \tag{34}$$

For the additive oracle, the only attainable syndromes are $0, U_1, \ldots, U_n$. The query is nonzero with probability $1 - 2^{-r}$, and conditional on that query no support equals it with probability $(1 - 2^{-r})^n$. Thus $R_{\mathcal{V}_1} = (1 - 2^{-r})^{n+1}$. The exact additive threshold is

$$n_{\mathcal{V}_1}(\epsilon) = \max\left\{0, \left\lceil \frac{\log \epsilon}{\log(1 - 2^{-r})} - 1 \right\rceil \right\}. \tag{35}$$

For fixed $\epsilon \in (0, 1)$, this is $\Theta(2^r)$, while $n \ge r + \lceil \log_2(1/\epsilon) \rceil$ suffices for Bayes. For $n < r$, the support rank is at most n, so $R^\star \ge 1 - 2^{n-r}$; fixed-risk Bayes sample complexity is also $\Omega(r)$. The result compares supports from a new task, independently of how many training tasks were used to construct a learner.

D.4 A finite-length interaction threshold

Define the accuracy-ϵ interaction complexity by

$$q_\epsilon(\Pi, n) = \min\{q \in \{0, \ldots, n\} : \Gamma_{>q} \le \epsilon\}. \tag{36}$$

We now analyse the full hierarchy in Section 4, rather than only its additive level. Let $T : \mathbb{F}_2^n \to \mathbb{F}_2^r$ be the random linear map whose columns are $U_1, \ldots, U_n$, and let $U \sim \mathrm{Unif}(\mathbb{F}_2^r)$ be an independent query. Define

$$C_q := \{c \in \mathbb{F}_2^n : \mathrm{wt}(c) \le q\}, \qquad B(n, q) := |C_q| = \sum_{j=0}^{q} \binom{n}{j}, \tag{37}$$

and count the order-q representations of the query by

$$N_q(U) := \sum_{c \in C_q} \mathbf{1}\{Tc = U\}, \qquad \mu_q := \mathbb{E} N_q(U) = \frac{B(n, q)}{2^r}. \tag{38}$$

Theorem 8 (Finite-length random-code threshold)**.** *For every n, r, and $q \in \{0, \dots, n\}$, the optimal order-q risk in the latent-code family obeys*

$$\left(1 - \frac{B(n,q)}{2^r}\right)_+ \le \inf_{g \in \mathcal{V}_q} R(g) \le \frac{1 - 2^{-r}}{B(n,q)2^{-r} + 1 - 2^{-r}} \le \frac{1}{1 + B(n,q)2^{-r}}. \tag{39}$$

Thus the transition occurs, up to finite-length constants, when the number of available coefficient vectors $B(n,q)$ crosses the number of possible syndromes 2^r.

Proof. By Theorem 6,

$$\inf_{g \in \mathcal{V}_q} R(g) = \Pr(U \notin T(C_q)) = \Pr(N_q(U) = 0). \tag{40}$$

The union bound gives $\Pr(N_q > 0) \le \mathbb{E}N_q = \mu_q$, proving the lower bound.

It remains to show that the $B(n,q)$ candidate representations have enough independence for a matching second-moment upper bound. For distinct $c, c' \in \mathbb{F}_2^n$, including the case where one is zero, the two linear constraints

$$Tc = U, \qquad Tc' = U \tag{41}$$

have rank two in the joint uniform variables (T, U). Consequently,

$$\Pr(Tc = U) = 2^{-r}, \qquad \Pr(Tc = U,\ Tc' = U) = 2^{-2r}. \tag{42}$$

The indicators in (38) are therefore pairwise independent, and

$$\mathbb{E}N_q^2 = \mu_q^2 + \mu_q(1 - 2^{-r}). \tag{43}$$

Paley–Zygmund yields

$$\Pr(N_q > 0) \ge \frac{(\mathbb{E}N_q)^2}{\mathbb{E}N_q^2} = \frac{\mu_q}{\mu_q + 1 - 2^{-r}}. \tag{44}$$

Taking the complement proves the middle upper bound in (39); the final simplification follows by elementary algebra. □

The minimum representation weight is a random-coset leader weight for the kernel code $\ker T$. This gives a standard coding-theoretic interpretation of the construction (MacWilliams & Sloane, 1977; Cohen et al., 1997); the finite-length bounds above are included to make the adaptation consequence self-contained.

Let

$$h_2(\delta) := -\delta \log_2 \delta - (1 - \delta) \log_2(1 - \delta) \tag{45}$$

denote binary entropy, with the usual continuous definitions at 0 and 1.

Theorem 9 (Entropy phase transition in interaction order)**.** *Fix $\alpha > 1$, let $n = \lceil \alpha r \rceil$, and define*

$$\delta_\star := h_2^{-1}(1/\alpha) \in (0, 1/2). \tag{46}$$

For any fixed $\delta \in (0, 1/2)$ and $q = \lfloor \delta n \rfloor$:

$$\delta < \delta_\star \quad \Longrightarrow \quad \inf_{g \in \mathcal{V}_q} R(g) \ge 1 - 2^{-[1 - \alpha h_2(\delta)]r + o(r)} \longrightarrow 1, \tag{47}$$

$$\delta > \delta_\star \quad \Longrightarrow \quad \inf_{g \in \mathcal{V}_q} R(g) \le 2^{-[\alpha h_2(\delta) - 1]r + o(r)} \longrightarrow 0. \tag{48}$$

Moreover, for every fixed $\epsilon \in (0, 1)$,

$$\frac{q_\epsilon(\Pi_r, n)}{n} \longrightarrow \delta_\star. \tag{49}$$

In particular, with only $O(r)$ supports, nontrivial Bayes accuracy requires interactions among $\Theta(r)$ support labels.

Proof. For fixed $\delta \in (0, 1/2)$, the Hamming-ball estimate gives

$$B(n, \lfloor \delta n \rfloor) = 2^{nh_2(\delta)+o(n)}. \tag{50}$$

Substitution into Theorem 8 proves (47) and (48). The unrestricted Bayes risk satisfies

$$R^\star \le 2^{r-n} = 2^{-(\alpha-1)r+O(1)} \longrightarrow 0 \tag{51}$$

by Theorem 6. Since $\Gamma_{>q} = \inf_{g \in \mathcal{V}_q} R(g) - R^\star$, every fixed $\delta < \delta_\star$ has tail tending to one, while every fixed $\delta > \delta_\star$ has tail tending to zero. Monotonicity in q squeezes q_ϵ/n to $\delta_\star$ for every $\epsilon \in (0, 1)$. □

Remark 10 (Critical window). Theorem 9 identifies the threshold location and gives exponential convergence at every fixed distance from it. The pairwise second-moment argument does not identify a limiting law in the critical window $B(n, q) \asymp 2^r$; no such law is needed for the adaptation lower bound.

D.5 Concentration of Bayes interaction mass

Theorem 11 (Bayes mass in linear-order conditional interactions)**.** *Fix $\alpha > 1$, let $n = \lceil \alpha r \rceil$, and put $\delta_\star = h_2^{-1}(1/\alpha)$. For any constants $0 < \delta_- < \delta_\star < \delta_+ < 1/2$,*

$$\sum_{q=\lceil \delta_- n \rceil}^{\lfloor \delta_+ n \rfloor} I_q \ge 1 - 2^{-c_- r+o(r)} - 2^{-c_+ r+o(r)}, \tag{52}$$

$$c_- := 1 - \alpha h_2(\delta_-) > 0, \qquad c_+ := \alpha h_2(\delta_+) - 1 > 0. \tag{53}$$

Consequently the latent-code Bayes target has $1 - o(1)$, and hence constant, squared mass in geometry-conditioned pure interactions of order $\Theta(r)$.

Proof. Theorem 6 gives $I_q = \Pr(k_{\min} = q)$. The union-bound half of Theorem 8 and the entropy estimate (50) imply

$$\Pr(k_{\min} < \delta_- n) \le 2^{-r} B(n, \lceil \delta_- n \rceil - 1) = 2^{-c_- r+o(r)}. \tag{54}$$

For the upper tail,

$$\Pr(k_{\min} > \lfloor \delta_+ n \rfloor) = \inf_{g \in \mathcal{V}_{\lfloor \delta_+ n \rfloor}} R(g) \le 2^{-c_+ r+o(r)} \tag{55}$$

by the second-moment half of the same theorem. The upper tail includes outside-span queries. Removing both tails proves (52). □

E A constructive contextual upper bound

The latent-code Bayes rule admits both an efficient algebraic implementation and an exact programmed attention implementation. The construction below is a capacity certificate with hard attention and exact arithmetic. Its mutable row memory differs from the learned Looped ICL architecture in Appendix A.3. General results on programmed looped Transformers provide related algorithmic constructions (Giannou et al., 2023).

Theorem 12 (Hard-attention elimination)**.** *For ambient binary dimension D and at most n supports, a recurrent hard-attention network with polynomial width, shared processor parameters, n mutable support tokens, one query token, and a null token computes the clean latent-code Bayes predictor exactly. It needs at most $\operatorname{rank}(X_{1:n}) \le r$ elimination rounds and a constant number of initialisation/readout stages. Each round consists of a constant number of attention and pointwise ReLU sublayers. Its parameters have polynomial bit complexity; no bounded-weight, finite-temperature, or learnability claim is made.*

Proof. Encode each support token by a row $v_i = X_i \in \mathbb{F}_2^D$ and a label bit $\ell_i \in \mathbb{F}_2$ with $Y_i = (-1)^{\ell_i}$. Store the query residual $v_{\mathrm{q}} = X_{\mathrm{q}}$ and an accumulator bit $a = 0$ separately. The null token stores a zero row and label bit zero.

Selecting and broadcasting a pivot. A hard-attention head gives support row v the scalar key $K(v) = \sum_{j=1}^{D} 2^{D-j} v_j$ and uses the constant query one at every receiving token. Query tokens are excluded from this source memory. The selected value is a lexicographically largest remaining row p and its label bit ℓ_p. A null source makes selection well defined when $n = 0$. If several rows maximise the key, they have identical geometry. They also have identical label bits because every updated row remains a valid linear combination of clean observed equations. Thus any tie choice, or an average of tied values, gives the same pivot. When every row is zero the pivot and its label bit are both zero.

Eliminating the pivot coordinate. If $p \neq 0$, let $j_\star$ be its first nonzero coordinate. Each support token sets $b_i = (v_i)_{j_\star}$ and performs the simultaneous update

$$v_i \leftarrow v_i \oplus b_i p, \qquad \ell_i \leftarrow \ell_i \oplus b_i \ell_p. \tag{56}$$

In particular, all copies of the pivot become zero rows with label bit zero. The query similarly uses $b_{\mathrm{q}} = (v_{\mathrm{q}})_{j_\star}$ and sets

$$v_{\mathrm{q}} \leftarrow v_{\mathrm{q}} \oplus b_{\mathrm{q}} p, \qquad a \leftarrow a \oplus b_{\mathrm{q}} \ell_p. \tag{57}$$

A zero pivot triggers no update. Every remaining row has zero in the chosen pivot coordinate. The map $v \mapsto v \oplus v_{j_\star} p$ has a one-dimensional kernel on the current row span, generated by p, so the remaining rank decreases by exactly one. After at most the initial rank many rounds, every support row is zero. Extra zero-pivot rounds are inert.

Correctness of the accumulated label. Elimination records the identity $X_{\mathrm{q}} = v_{\mathrm{q}} \oplus \bigoplus_t b_{\mathrm{q},t} p_t$. The pivots form a basis of the original support span; the final query residual is zero exactly when the query lies in that span. Every pivot label equals the latent character evaluated on its row, since (56) preserves each observed equation. Hence, if the query residual is zero, $(-1)^a$ is its query label. If it is nonzero, the Bayes mean is zero by Theorem 6. A zero query retains $a = 0$ and returns $+1$. This proves the required readout.

Realisation by pointwise networks. All state variables are bits. AND and XOR have exact ReLU implementations, for example $u \wedge v = \mathrm{ReLU}(u + v - 1)$ and $u \oplus v = u + v - 2\,\mathrm{ReLU}(u + v - 1)$ on $\{0, 1\}^2$. The first-nonzero indicators are $e_j = \mathrm{ReLU}(p_j - \sum_{h<j} p_h)$. There is at most one nonzero e_j, so $b_i = \sum_j (e_j \wedge (v_i)_j)$ is again a bit. These gates implement (56)–(57) in constant depth and polynomial width. Token-type bits gate support and query updates, and residual channels retain their states. A final zero-test and sign conversion produce $\mathbf{1}\{v_{\mathrm{q}} = 0\}(1 - 2a)$. All rounds reuse the same operations. Although the selection keys can have magnitude exponential in D, their integer representation needs only $O(D)$ bits. This completes the exact hard-attention construction. □

The construction separates architectural realisability from the route selection that outer training must discover. The minimum-cardinality route used to define $k_{\min}$ is unnecessary for prediction: any valid elimination route yields the same clean label. Neither this construction nor the local-multiplication theorem proves that endpoint-trained soft attention will learn the elimination algorithm.

F How useful evidence is exposed to gradients

This appendix gives a restricted analytic example, separate from the empirical models. Its single geometry-only block, fixed gain with $\varrho < 1$, and bounded parameter region are essential assumptions, not properties established for multi-layer Looped ICL or Set ICL. The positive local-product control does not cover learning the geometry-dependent pivot policy.

For a fixed geometry, let $m_H(Y) = \mathbb{E}[Y_{\mathrm{q}} \mid H, Y]$ and $J_{\theta,H}$ be the parameter-to-function Jacobian of $F_\theta(H, \cdot)$. We introduce gradient access as a diagnostic for this paper: define conditional target access and its energy-normalised average by

$$A_H(\theta) = \left\| \mathbb{E}_{Y|H}[m_H(Y) \nabla_\theta F_\theta(H, Y)] \right\|_2^2, \qquad \bar{\mathfrak{a}}(\theta) = \frac{\mathbb{E}_H A_H(\theta)}{\mathbb{E}_H \|m_H\|_2^2}. \tag{58}$$

The numerator measures alignment with a specified target, not total gradient magnitude. Averaging conditional squared norms upper-bounds the squared norm of the global target-aligned gradient, by Jensen's inequality.

Consider p geometry-only attention heads with label averages $s_j = \sum_i a_{ji}(H;\omega)Y_i$, $a_{ji} \geq 0$, $\sum_i a_{ji} = 1$, and

$$F_\theta(H,Y) = b_0 + \sum_{j=1}^{p} c_j\, \sigma_j\big(\beta_j(H;\omega) + \gamma s_j\big). \tag{59}$$

The value gain $\gamma > 0$ is fixed, the features are analytic, and the final readout is initialised at $b_0 = c = 0$.

Theorem 13 (Restricted endpoint access and a local positive control)**.** *Suppose the activation Taylor coefficients and their first derivatives are bounded by MR^{-d}, with fixed $\varrho = \gamma/R < 1$, and the widths, parameter scales, and geometry derivatives are polynomially bounded in r throughout a parameter region. Then, for an in-span latent-code query,*

$$A_H(\theta) \leq Q(r)\varrho^{2k_{\min}(H)}, \qquad \sup_\theta \bar{\mathfrak{a}}(\theta) \leq \exp[-\Omega(r)] \quad \textit{when } n = \lceil \alpha r \rceil,\ \alpha > 1, \tag{60}$$

where Q is polynomial. In contrast, for independent uniform signs Z, Y, the local target ZY and feature $h(Z,Y) = \mathrm{GELU}(\gamma(Z+Y)/2)$ have

$$\mathbb{E}[ZYh(Z,Y)] = \frac{\gamma[2\Phi(\gamma) - 1]}{4} > 0. \tag{61}$$

A zero readout on this feature therefore has constant local target access, independent of latent rank.

The mechanism is conditional character cancellation: every Taylor term below $k_{\min}$ is orthogonal to the endpoint target; analyticity controls the remaining tail. Appendix G states the full parameter bounds and proofs. A teacher-forced running product exposes the pair (Z, Y) directly, making the same multiplication rule available at every step.

An initial access bound is distinct from a training-time result. If global target alignment remains exponentially small and the Jacobian norm remains polynomially bounded in a sufficiently large parameter neighbourhood, zero-output population gradient flow has an exponential persistence time. The discrete-gradient analogue additionally requires a nonexpansive zero-target update (Appendix G.2). These statements concern the specified geometry-only analytic model. Mutable recurrent attention may change its label dependence and leave this regime. The local result exposes label multiplication; learning pivot selection and autonomous rollout remains outside its guarantee.

G GRADIENT-ACCESSIBILITY THEORY

The following results use the conditional label law, rather than a uniform Boolean-cube surrogate. They concern the analytic geometry-only cross-attention block specified below. All parameter-gradient norms use its stated Euclidean parameter coordinates.

For a fixed geometry H, let $L^2_H = L^2(Y \mid H)$, let $m_H(Y) = \mathbb{E}[Y_q \mid H, Y]$, and let $J_{\theta,H} : \mathbb{R}^P \to L^2_H$ be the derivative of $\theta \mapsto F_\theta(H,\cdot)$. We measure conditional target access by

$$A_H(\theta) := \|J^*_{\theta,H} m_H\|_2^2, \qquad \bar{\mathfrak{a}}(\theta) := \frac{\mathbb{E}_H A_H(\theta)}{\mathbb{E}_H \|m_H\|^2_{L^2_H}}. \tag{62}$$

The numerator is the squared norm of the parameter gradient of the conditional endpoint correlation, before averaging gradients across geometries. It is thus an upper bound on the access available to shared parameters.

Proposition 14 (Conditional access dominates global gradient alignment)**.** *Let $J_\theta : \mathbb{R}^P \to L^2(H,Y)$ be the global Jacobian and suppose the quantities in* (62) *are finite. Then*

$$\frac{\|J^*_\theta m\|_2^2}{\|m\|_2^2} \leq \bar{\mathfrak{a}}(\theta). \tag{63}$$

Proof. Disintegration of the population inner product gives $J^*_\theta m = \mathbb{E}_H[J^*_{\theta,H} m_H]$. Jensen's inequality therefore yields $\|J^*_\theta m\|_2^2 \leq \mathbb{E}_H \|J^*_{\theta,H} m_H\|_2^2$, while $\|m\|_2^2 = \mathbb{E}_H \|m_H\|^2_{L^2_H}$. □

The next elementary lemma is the source of the small endpoint gradient. It uses the conditional code C_H and syndrome map T_H from (30).

Proposition 15 (Low-degree cancellation on a conditional label code)**.** *Fix H and a query in the support span. Choose c with $T_H c = U_q$ and let $k = q_H(c) = k_{\min}(H)$. If $P_d(Y_1, \dots, Y_n)$ is an ordinary polynomial of total degree at most $d < k$, then*

$$\mathbb{E}_{Y \sim \mathrm{Unif}(C_H)}[\chi_c(Y) P_d(Y)] = 0. \tag{64}$$

Proof. On sign inputs, every monomial reduces to a multilinear character χ_a with $\mathrm{wt}(a) \le d$. Character orthogonality on C_H says $\mathbb{E}\chi_{a+c} = 1$ exactly when $a + c \in \ker T_H$, and is zero otherwise. The first case would place a in $c + \ker T_H$ with weight below the coset leader weight k, a contradiction. Linearity proves the claim. □

G.1 A SPECIFIED CROSS-ATTENTION PARAMETERISATION

Consider p geometry-only attention heads followed by analytic scalar features and a linear readout:

$$s_j(H, Y; \omega) := \sum_{i=1}^{n} a_{ji}(H; \omega) Y_i, \qquad a_{ji} \ge 0, \quad \sum_i a_{ji} = 1, \tag{65}$$

$$F_\theta(H, Y) := b_0 + \sum_{j=1}^{p} c_j \sigma_j(\beta_j(H; \omega) + \gamma s_j(H, Y; \omega)). \tag{66}$$

This is a one-query multihead cross-attention block: query–key scores and therefore a_{ji} are functions of H alone, scalar value channels contain Y_i with gain γ, the σ_j are pointwise FFN features, and (b_0, c) is the final readout. We initialise $b_0 = c_1 = \cdots = c_p = 0$.

Let $\mathcal{B}$ be a parameter region containing this initialisation. Assume throughout $\mathcal{B}$ that $|c_j| \le C_c$, that ω has P_ω scalar coordinates, and that for every such coordinate ω_a,

$$|\partial_a \beta_j| + \sum_i |\partial_a a_{ji}| \le G. \tag{67}$$

Assume also that for some $M, R > 0$, all $d \ge 0$, and $\nu \in \{0, 1\}$,

$$\frac{|\sigma_j^{(d+\nu)}(\beta_j)|}{d!} \le M R^{-d}, \qquad \varrho := \gamma / R < 1. \tag{68}$$

The bounds are uniform in $(H, \theta) \in \mathcal{B}$. A softmax geometry attention satisfies (67) whenever its score derivatives are bounded, since $\partial a_i = a_i(\partial \ell_i - \sum_l a_l \partial \ell_l)$. Analytic activations such as GELU, or activations analytic on a sufficiently large complex neighbourhood of the relevant preactivation interval, admit such bounds on compact regions with a fixed suitable radius; the associated constant M must still obey the stated polynomial scale control. The theorem below is coordinate-scale dependent, as any parameter-gradient statement must be.

Theorem 16 (Exponentially small endpoint access for analytic attention)**.** *For the model* (66)*, condition on a latent-code geometry whose query is in the support span and has $k = k_{\min}(H)$. Uniformly over $\theta \in \mathcal{B}$,*

$$\| J^*_{\theta, H} m_H \|_2^2 \le Q(r) \varrho^{2k}, \tag{69}$$

where one valid bound is

$$Q(r) := 1 + \frac{M^2}{(1 - \varrho)^2} \left[p + P_\omega p^2 C_c^2 G^2 (1 + R)^2 \right]. \tag{70}$$

The separate leading 1 covers the bias coordinate when $k = 0$. If $p, P_\omega, C_c, G, M, R, (1 - \varrho)^{-1}$ are polynomially bounded in r and $\varrho < 1$ is bounded away from one, then for $n = \lceil \alpha r \rceil$ and every fixed $0 < \delta < \delta_\star$,

$$\sup_{\theta \in \mathcal{B}} \overline{\mathfrak{a}}(\theta) \le \mathrm{poly}(r) \left[\varrho^{2\delta n} + 2^{-[1 - \alpha h_2(\delta)] r + o(r)} \right] = \exp[-\Omega(r)]. \tag{71}$$

In particular the conclusion holds at the specified zero-readout initialisation, and Proposition 14 implies the same exponential upper bound for normalised global endpoint gradient alignment.

Proof. For a readout coordinate, expand at the geometry-dependent preactivation:

$$\sigma_j(\beta_j + \gamma s_j) = \sum_{d=0}^{\infty} \frac{\sigma_j^{(d)}(\beta_j)}{d!} \gamma^d s_j^d. \tag{72}$$

Because s_j^d is a polynomial of label degree at most d, Proposition 15 eliminates every term with $d < k$. Moreover $|s_j| \le 1$, so absolute convergence and (68) give

$$|\langle m_H, \partial_{c_j} F_\theta \rangle_H| \le M \sum_{d \ge k} \varrho^d = \frac{M \varrho^k}{1 - \varrho}. \tag{73}$$

For a geometry parameter, write $t_{ja} = \sum_i (\partial_a a_{ji}) Y_i$. Differentiation gives

$$\partial_a F_\theta = \sum_j c_j \sigma_j'(\beta_j + \gamma s_j)(\partial_a \beta_j + \gamma t_{ja}). \tag{74}$$

In the first term, cancellation leaves only Taylor orders $d \ge k$. In the second, $t_{ja} s_j^d$ has degree at most $d+1$, so cancellation leaves only $d \ge k-1$. Since $|t_{ja}| \le G$, the two geometric tails are bounded by

$$\frac{MG\varrho^k}{1-\varrho} \quad \text{and} \quad \frac{MGR\varrho^k}{1-\varrho}, \tag{75}$$

respectively. Thus each geometry-coordinate correlation is at most $pC_cMG(1+R)\varrho^k/(1-\varrho)$. Squaring and summing the p readout and P_ω geometry coordinates, with the constant bias treated separately, proves (69)–(70).

For the population statement, split according to $k_{\min} \ge \delta n$. The low-weight event has probability at most

$$2^{-r} B(n, \lceil \delta n \rceil - 1) = 2^{-[1 - \alpha h_2(\delta)]r + o(r)}. \tag{76}$$

On its complement (69) contributes at most $Q(r)\varrho^{2\delta n}$; outside the support span $m_H = 0$. Finally, $\mathbb{E}_H \|m_H\|^2 = \Pr(k_{\min} < \infty) \ge 1 - 2^{r-n}$, because a union bound over nonzero dual vectors shows that the random support matrix fails to have rank r with probability below 2^{r-n}. Division proves (71). □

Remark 17 (Why the contraction condition is substantive). The ratio $\varrho < 1$ says that, throughout the stated analytic region, the label perturbation of a scalar feature remains strictly inside a uniform analytic radius. If gains, widths, readout scales, or activation derivatives grow exponentially with r, the polynomial prefactor conclusion fails. Likewise, label-dependent query/key features can invalidate the low-degree Taylor cancellation. These are mechanisms by which training may leave the proved low-alignment regime.

G.2 Persistence and training-time consequences

The next result says that a uniformly small target-aligned gradient cannot suddenly become useful before the trajectory has had enough time to leave the region on which the bound was proved.

Theorem 18 (Persistence of a nonlinear low-alignment region)**.** *Let $\mathcal{E}(\theta) = \frac{1}{2}\|F_\theta - m\|_2^2$ and consider population gradient flow*

$$\dot{\theta}_t = J_{\theta_t}^*(m - F_{\theta_t}), \qquad F_{\theta_0} = 0. \tag{77}$$

Let $B, \varepsilon, R_0 > 0$. Suppose that throughout the parameter ball $\mathbb{B}(\theta_0, R_0)$,

$$\|J_\theta\|_{\mathrm{op}} \le B, \qquad \|J_\theta^* m\|_2 \le \varepsilon. \tag{78}$$

Then the trajectory cannot leave the ball before

$$T_{R_0} := \frac{\sqrt{1 + 2B^2 R_0/\varepsilon} - 1}{B^2}, \tag{79}$$

and for every $t \le T_{R_0}$,

$$\|F_{\theta_t}\|_2 \le B\varepsilon t, \tag{80}$$

$$\|\theta_t - \theta_0\|_2 \le \varepsilon t + \tfrac{1}{2} B^2 \varepsilon t^2. \tag{81}$$

Hence if B is polynomially bounded, R_0 is inverse-polynomially bounded below, and $\varepsilon = \exp[-\Omega(r)]$, no constant-norm endpoint predictor is reached for $\exp[\Omega(r)]$ population time unless training exits the region where (78) *holds.*

Proof. While the trajectory is in the ball,

$$\frac{d}{dt}\frac{1}{2}\|F_{\theta_t}\|_2^2 = \langle J_{\theta_t}^* F_{\theta_t}, J_{\theta_t}^* m - J_{\theta_t}^* F_{\theta_t}\rangle \\ \le B\varepsilon\|F_{\theta_t}\|_2. \tag{82}$$

Thus $d\|F_{\theta_t}\|_2/dt \le B\varepsilon$ (by continuity at zero), proving (80). The parameter speed obeys

$$\|\dot\theta_t\|_2 \le \varepsilon + B\|F_{\theta_t}\|_2 \le \varepsilon + B^2\varepsilon t. \tag{83}$$

Integration proves (81). The right side first equals R_0 at (79); a standard first-exit argument therefore proves the persistence claim. The final statement follows by substitution. □

Theorem 16 bounds squared access. Thus its uniform parameter-gradient bound has $\varepsilon = \sqrt{\mathbb{E}_H A_H} = \exp[-\Omega(r)]$ after absorbing polynomial factors and using Proposition 14. Combined with Theorem 18, it proves an exponential population-flow plateau for every polynomially controlled ball contained in $\mathcal{B}$.

For literal gradient descent, a weak alignment bound alone does not control the label-independent self-loss update. The following condition cleanly states the additional dynamical assumption needed.

Corollary 19 (Discrete endpoint-gradient lower bound in a nonexpansive basin)**.** *In the setting of Theorem 18, define the zero-target self-loss map*

$$G_\eta(\theta) := \theta - \eta J_\theta^* F_\theta. \tag{84}$$

Suppose, throughout $\mathbb{B}(\theta_0, R_0)$*,*

$$\|G_\eta(\theta) - \theta_0\|_2 \le \|\theta - \theta_0\|_2, \quad \|J_\theta^* m\|_2 \le \varepsilon, \quad \|J_\theta\|_{\mathrm{op}} \le B. \tag{85}$$

Then endpoint gradient descent satisfies, as long as it remains in the ball,

$$\|\theta_t - \theta_0\|_2 \le \eta\varepsilon t, \qquad \|F_{\theta_t}\|_2 \le B\eta\varepsilon t. \tag{86}$$

Consequently no predictor of norm at least ζ *is reached before*

$$t \ge \min\left\{\frac{R_0}{\eta\varepsilon}, \frac{\zeta}{B\eta\varepsilon}\right\}. \tag{87}$$

For inverse-polynomial step size and the analytic-attention access bound, this is $\exp[\Omega(r)]$ *steps. Condition* (85) *holds for the corresponding stable linearized or frozen-trunk readout dynamics; it is not established here for a fully nonlinear mutable Transformer.*

Proof. The endpoint update decomposes as

$$\theta_{t+1} = G_\eta(\theta_t) + \eta J_{\theta_t}^* m. \tag{88}$$

The triangle inequality and (85) inductively give the parameter bound in (86). Since $F_{\theta_0} = 0$, the Jacobian bound makes the parameter-to-function map B-Lipschitz on the ball, proving the output bound. The first quantity in (87) prevents exit and the second prevents norm ζ. □

G.3 A CONSTANT-ACCESS INTERMEDIATE OBJECTIVE

The running-product trace has a matching positive result at the same style of analytic readout. Let Z, Y be independent uniform signs, let the attention average their value channels, and define $s = (Z + Y)/2$. The local target is $\mu(Z, Y) = ZY$.

Proposition 20 (Constant gradient access for a supervised product step)**.** *At a zero readout multiplying the tangent feature* $h_{b,\gamma}(Z, Y) = \sigma(b + \gamma s)$*, the target-aligned gradient is*

$$\kappa_\gamma(b) := \mathbb{E}[ZY h_{b,\gamma}(Z, Y)] = \frac{\sigma(b+\gamma) + \sigma(b-\gamma) - 2\sigma(b)}{4}. \tag{89}$$

Whenever $\kappa_\gamma(b) \ne 0$*, the local objective has access at least* $\kappa_\gamma(b)^2$*, a constant independent of* r*. For GELU and* $b = 0$*,*

$$\kappa_\gamma(0) = \frac{\gamma[2\Phi(\gamma) - 1]}{4} > 0 \quad (\gamma > 0). \tag{90}$$

Alternatively, exposing the centered quadratic feature $h(Z, Y) = 2s^2 - 1$ *gives* $h = ZY$ *exactly and unit access.*

Proof. Enumerating the four pairs (Z, Y) gives (89). Since $\mathrm{GELU}(x) = x\Phi(x)$ and $\Phi(-x) = 1 - \Phi(x)$, its specialization is (90). Finally $s^2 = (1 + ZY)/2$, proving the exact quadratic identity. □

Theorems 16 and 18 link a uniform conditional-access bound to a population-flow plateau under the stated Jacobian and neighbourhood bounds. Corollary 19 adds the condition needed for discrete updates. Proposition 20 exposes a shared multiplication transition with dimension-free access. This is the scope of the endpoint-versus-local comparison: it establishes neither polynomial-time learning of pivot selection nor successful autonomous Gaussian-elimination rollout for the trained Looped Transformer.