MAPPING THE SOLUTION-SET STRUCTURE OF GROKKING ON LATIN SQUARES

Anonymous authors
Paper under double-blind review

ABSTRACT

Grokking is a delayed transition from memorisation to generalisation. A network first fits observed examples while remaining near chance on held-out inputs, then generalises abruptly. Prior work has focused mainly on this delay, finding that the network's main loss becomes weight decay after the fitting phase. We name the solution to this optimization problem as the minimum-norm endpoint, and study which tasks make all minimum-norm endpoints generalise from partial observations. Specifically, we model pairwise input tasks as edge colourings of $K_{q,q}$, with Latin colourings as our structured family, and classify each observed instance as all-good, mixed, or no-good by its endpoint set. Furthermore, we prove two contrasting results: (1) Finite-abelian-group families are asymptotically all-good from partial observations; (2) Under a uniform random Latin square, all-good tasks become asymptotically rare for observation-based learners, leaving mixed or no-good behaviour. Moreover, experiments show that relabellings of addition generalise despite their dense polynomial representations. Finally, on tasks of composite order, finite-training results suggest that predictors learn a coarse part of the output before the full label.

1 INTRODUCTION

Grokking is delayed generalisation: a network fits its training examples, remains near chance on held-out inputs for a long period, and then generalises abruptly (Power et al., 2022). The classical instance is modular addition, where the network observes some pairs (x, y) with their sums $x + y \bmod q$ and must infer the rest. Such a learning curve combines two things, the quality of the predictor that training eventually reaches and the time taken to reach it, and existing work examines how learned representations, regularisation, and the training data shape this behaviour (Nanda et al., 2023; Mohamadi et al., 2024; Gu et al., 2025). However, a run that generalises shows that there exsits one solution does; it does not show that every solution a training could have selected would. We isolate the generalisation behaviour and ask it at the level of the task. Defining the minimum-norm endpoint as the solution to the weight decay-induced optimization problem, our central question is therefore: *which task structures make the entire minimum-norm endpoint set generalise?*

Answering it needs modelling of the task and endpoint set. A pairwise input task assigns an output to every pair of inputs, and we encode it as an *edge colouring* of the complete bipartite graph $K_{q,q}$: the two input alphabets are the two vertex classes, each pair (u, v) is an edge, which we also call a cell, and the output is its colour. Written out, the colouring is a $q \times q$ table, and at prime-power order it is a polynomial over $\mathbb{F}_q$; table, graph, and polynomial are three coordinates on one object. The colouring is *Latin* when it is proper, so that no vertex sees a colour twice; in table coordinates every symbol appears exactly once in each row and column, so the Latin colourings are the Latin squares of combinatorics. Every group multiplication table is one, and the family therefore contains modular addition together with its natural competitors. Permuting the vertices within either class, permuting the colours, or swapping the two classes only changes how the table is written; we call such a move a *relabelling*, and the orbit of a colouring under relabellings is our task object. Training observes a uniformly random subset of the cells of fixed size.

The network reads an edge through its two endpoints, with one embedding table per vertex class, a hidden ReLU layer, and a head over the colours. For the endpoint of training we follow implicit-bias theory, which motivates the minimum-norm interpolator as the idealised late-training object (Soudry

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.

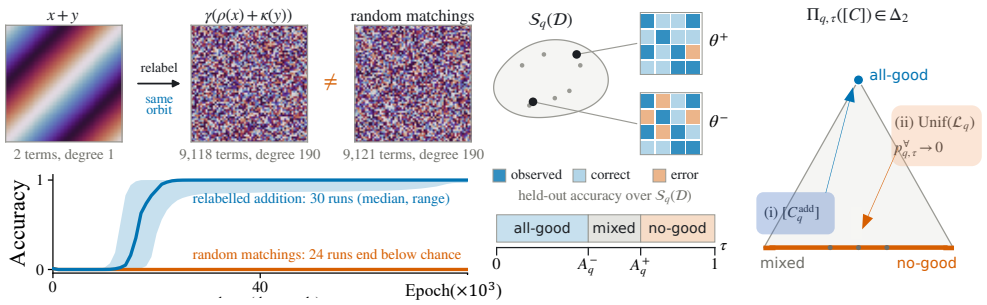


Figure 1: **Coordinates, endpoint sets, and the two poles.** Left: $x + y \bmod 97$; a relabelling of it, 9,118 terms and degree 190; and a random-matching colouring outside the orbit. Middle: the minimum-norm interpolators $\mathcal{S}_q(\mathcal{D})$ agree on the observed cells, may differ elsewhere; their extremes $A_q^- \le A_q^+$ cut τ into all-good, mixed, no-good. Right: each orbit maps to a simplex point; abelian group families sit at the all-good vertex (cyclic addition illustrated), while the uniform Latin ensemble sits on the opposite edge, where the mixed/no-good split is open.

et al., 2018; Lyu & Li, 2020; Neyshabur et al., 2015). A partially observed colouring, however, has many minimum-norm interpolators: all meet the margins on the observed cells and are free to disagree on the rest, and a handful of training runs yields a few selected predictors without certifying that they belong to this set or how the set behaves as a whole. We therefore take the whole set, the *endpoint set*, as the object of study, and describe it by what holds across it at a fixed held-out threshold τ: an observed instance is all-good if every minimum-norm interpolator generalises, no-good if none does, and mixed if both kinds coexist. Averaged over the random observations, the three probabilities form the task's *phase profile*, a point of a triangle (Figure 1, middle and right).

Relabelling induces a norm-preserving bijection between endpoint sets and preserves the law of the observations, so the phase profile is well defined on orbits: it is a map from colouring orbits to the triangle. The network realises the bijection by permuting the rows of its embedding tables and of its head. This separates task structure from its polynomial description. Take $x + y \bmod 97$ and relabel it. The endpoint set is transported by the bijection, but the polynomial coordinate becomes a polynomial over $\mathbb{F}_{97}$ with 9,118 nonzero terms and total degree 190, with essentially the same support density and the same degree as a proper colouring assembled from random perfect matchings, which lies outside the orbit. Networks trained on the relabelled colouring memorise and then generalise as they do on $x + y$; networks trained on the random-matching colouring end below chance (Figure 1, left). Relabelling preserves the phase while changing polynomial sparsity and degree, and in this comparison those coordinate statistics also fail to distinguish the trained predictors' outcomes. This is why we work with orbits.

Our results place two structures at opposite poles of the triangle. The all-good pole contains *every finite abelian group family*: at any fixed positive observation fraction, the worst endpoint accuracy converges to one in probability as the order grows. Cyclic addition is one instance, and non-isomorphic abelian groups supply distinct all-good orbits. At the opposite pole, a uniformly random Latin colouring defeats every learner that reads only the observations, at any fixed positive accuracy; consequently the colourings with a non-negligible all-good component have vanishing uniform mass. Generalisation is thus guaranteed throughout a structured family yet rare in the ambient space.

The core of our theory is a lemma of *carrier* criterion: a network that scores the full grid with controlled margins at a known norm cost. It bounds the cost of every minimum-norm endpoint on every mask, and a finite atlas fixed before the observations counts the harmful low-norm alternatives, whose number is too small for any of them to evade a random mask. The criterion does not depend on the colouring, and abelian characters provide carriers uniformly across group structures. Therefore, we have proven that satisfying this lemma is a sufficient condition for all good. We conjecture that the carrier criterion is also necessary: every asymptotically all-good Latin family admits a full-grid carrier in the same norm-and-margin regime.

Experiments at composite order 64 compare the guaranteed families with non-abelian and non-group candidates and show that coarse output factors are recovered before exact labels. These trained predictors probe structure; they do not measure the phase of the whole endpoint set. Whether a typical random colouring has some good endpoint, and whether any family carries non-vanishing mixed mass, remain open.

Contributions.

- A task-level formulation: Latin colourings of $K_{q,q}$ compared up to relabelling, each observed instance judged by its whole minimum-norm solution set.
- Two poles: every finite abelian group family is asymptotically all-good, while the uniform Latin ensemble defeats every learner at any fixed positive held-out accuracy.
- A colouring-independent carrier criterion for all-goodness, together with the conjecture that this criterion is necessary.
- Experiments in the same model class: relabelled addition groks like addition while looking like noise, and at composite order structured tasks recover coarse output factors before exact labels.

Paper Organization. Section 2 reviews prior work, Section 3 defines the model, Section 4 states and proves the two poles and the abelian extension, Section 5 reports the composite-order experiments and states the conjecture, and Section 6 concludes. Experiments that speak directly to the theory follow the theory itself.

2 Related Work

Grokking is reported in non-neural models (Mallinar et al., 2024), deep networks (Fan et al., 2024; Manir & Rupa, 2026), in-context and multi-hop reasoning (He et al., 2024; Wang et al., 2024; Abramov et al., 2025), and LLM pretraining (Li et al., 2025); its onset needs a critical data fraction (Zhu et al., 2024; Huang et al., 2024) and its delay can be shortened (Thilak et al., 2022; Lee et al., 2024; Park et al., 2024; Kunin et al., 2024; Xu et al., 2025). Related approaches include representation learning, norm and optimisation bias, and circuit competition. *Representation-level* work studies Fourier circuits and progress measures (Nanda et al., 2023; Gromov, 2023; Furuta et al., 2024; Barak et al., 2022), spectral dynamics (Zhou et al., 2024; Yunis et al., 2024), embedding bottlenecks (AlQuabeh et al., 2025), architectural bias (Singh et al., 2026; Yıldırım, 2026), task symmetry (Hwang & Park, 2026), embedding uniformity coupled to the training distribution (Gu et al., 2025), feature-emergence scaling laws (Tian, 2025), and topological signatures (Tang et al., 2026). *Norm- and optimisation-based* accounts appeal to implicit bias toward max-margin, minimum-norm endpoints (Soudry et al., 2018; Lyu & Li, 2020; Neyshabur et al., 2015; Lyu et al., 2024; Mohamadi et al., 2024; Humayun et al., 2024), to norm minimisation on the zero-loss manifold (Muşat, 2025; Tikeng Notsawo et al., 2025), to the edge of linear separability (Beck et al., 2025), to numerical stability (Prieto et al., 2025), sharpness (Miller et al., 2024), Jacobian alignment (Walker et al., 2025), and quantitative delay laws (Truong et al., 2026; Tan et al., 2026; Ersoy & Wiesner, 2026). *Competition* accounts pit a slow generalising circuit against a fast memorising one (Varma et al., 2023; Merrill et al., 2023), linking grokking to double descent and emergent abilities (Davies et al., 2023; Huang et al., 2024; Liu et al., 2022), with phase-transition, statistical, basin-selection, and capacity variants (Clauw et al., 2024; Carvalho et al., 2025; Cullen et al., 2026; Song & Ye, 2026; Luo et al., 2026). Provable results include generalisation guarantees for bounded-norm quadratic networks on modular addition (Mohamadi et al., 2024) and delay bounds in ridge regression (Xu et al., 2026). Gu et al. (2025) separately analyse the accuracy ceiling and embedding dynamics; Appendix E.2 discusses their ceiling and transition-time measurements. Our setting fixes a ReLU architecture and a uniform mask law, compares Latin colourings up to relabelling, and quantifies over every minimum-norm response. These choices specify the scope of our endpoint guarantees.

3 Latin-Square Tasks and the Endpoint Set

This section builds the objects the introduction used informally: the task and its observed part, the endpoint set, its three phases, and the relabelling invariance that makes the phase a property of the task rather than of its presentation.

3.1 Tasks, masks, and three coordinates

Let $q \geq 2$, $[q] = \{0, \ldots, q-1\}$, $X_q = [q]^2$, and $N = q^2$. A pairwise classification task is a colouring $C : X_q \to [q]$; the set of all colourings is $\mathcal{C}_q$. A colouring is *Latin* when every symbol occurs exactly once in each row and each column; $\mathcal{L}_q \subset \mathcal{C}_q$ denotes the Latin colourings of order q. For every q,

modular addition $C_q^{\mathrm{add}}(u,v) = u + v \pmod q$ is Latin and is the cyclic prototype of the abelian all-good family proved below.

Fix $\alpha \in (0,1)$. The observed cells form a uniformly random mask of fixed size,

$$\Omega \sim \nu_q := \mathrm{Unif}\big\{\Omega' \subseteq X_q : |\Omega'| = \lfloor \alpha N \rfloor\big\}, \qquad T = X_q \setminus \Omega; \tag{1}$$

T is the held-out set and $\mathcal{D}(C,\Omega) = \{(u,v,C(u,v)) : (u,v) \in \Omega\}$ the dataset. All asymptotic statements hold at fixed α.

Proposition 3.1 (Two coordinates, and a third at prime-power order)**.** *For every $q \geq 2$, a label table $C \in \mathcal{C}_q$ and a q-edge-colouring of $K_{q,q}$ (colour $C(u,v)$ on the edge U_uV_v) determine each other. When q is a prime power and $[q]$ is identified with $\mathbb{F}_q$, both are also equivalent to the unique reduced polynomial $P_C \in \mathbb{F}_q[X,Y]$ with $\deg_X P_C, \deg_Y P_C \leq q-1$ and $P_C(u,v) = C(u,v)$. C is Latin if and only if the edge colouring is proper, i.e. each colour class is a perfect matching.*

3.2 The minimum-norm endpoint set

The model is a bias-free one-hidden-layer ReLU network with a separate embedding table for each input, $E_U, E_V : [q] \to \mathbb{R}^{d_q}$, a hidden layer $W \in \mathbb{R}^{h_q \times d_q}$, and a head $A \in \mathbb{R}^{q \times h_q}$, where $d_q = 2q+1$ and $h_q = q^2$; the width allows one hidden unit per cell, which is what feasibility needs, and the embedding dimension exceeds the rank of any reduced representation. Writing $\theta = (E_U, E_V, W, A)$,

$$F_\theta(u,v) = A\ \mathrm{ReLU}\big(W\,(E_U(u) + E_V(v))\big) \in \mathbb{R}^q,\ \ \|\theta\|^2 = \|E_U\|_F^2 + \|E_V\|_F^2 + \|W\|_F^2 + \|A\|_F^2. \tag{2}$$

As a function class the pair embedding contains the common design that shares one table E and concatenates $[E(u); E(v)]$, since $W[E(u);E(v)] = W_1E(u) + W_2E(v)$; the norms differ, so the endpoint sets need not coincide. The *endpoint set* of an observed dataset is the set of global minimisers of the parameter norm under unit margins on the observed cells,

$$\mathcal{S}_q(\mathcal{D}) = \arg\min_\theta\ \tfrac{1}{2}\|\theta\|^2 \quad \text{s.t.} \quad F_{\theta,y}(u,v)\ \geq\ 1 + \max_{a \neq y} F_{\theta,a}(u,v) \qquad \big((u,v,y) \in \mathcal{D}\big). \tag{3}$$

Max-margin and implicit-bias results for homogeneous networks motivate this as the idealised late-training object (Soudry et al., 2018; Lyu & Li, 2020; Neyshabur et al., 2015). $\mathcal{S}_q(\mathcal{D})$ is a set: all its members meet the margins on Ω, and they may predict differently on T. Appendix B.2 treats the extension to L_2-regularised biases.

Three of the four parameter blocks can be folded into one matrix, which is the form the proofs work with. Assemble $\Phi = [E_U(0), \dots, E_U(q-1), E_V(0), \dots, E_V(q-1)] \in \mathbb{R}^{d_q \times 2q}$, put $Z = W\Phi$ and $\xi_{(u,v)} = e_u + e_{q+v} \in \mathbb{R}^{2q}$, so that $F_\theta(x) = A\,\mathrm{ReLU}(Z\xi_x)$. Call (A,Z) a *reduced representation* and define

$$\begin{aligned} J(F) &= \inf\big\{\|A\|_F\,\|Z\|_* :\ F = A\,\mathrm{ReLU}(Z\xi.),\ \text{any finite width}\big\}, \\ J_* &= \min\big\{\|A\|_F\,\|Z\|_* :\ Z \in \mathbb{R}^{h_q \times 2q},\ A\,\mathrm{ReLU}(Z\xi.)\ \text{meets the margins of (3)}\big\}. \end{aligned} \tag{4}$$

Lemma 3.2 (Reduced normalisation)**.** *The optimal value of* (3) *is $V_* = \frac{3}{2}J_*^{2/3}$, and every global minimiser is balanced: $\|A\|_F^2 = \|W\|_F^2 = \|\Phi\|_F^2 = J_*^{2/3}$ and $\|A\|_F\|W\Phi\|_* = J_*$.*

In words: the nuclear norm of Z is the cost of its cheapest factorisation into a hidden layer and an embedding, and the scale freedom of the ReLU then balances the head against the rest, so an endpoint is simply a score table whose product $\|A\|_F\|Z\|_*$ is as small as the observations allow. This product is the one quantity the main proof tracks.

3.3 Phases of the endpoint set and the phase profile

Fix C and Ω and write $\mathcal{S}_q = \mathcal{S}_q(\mathcal{D}(C,\Omega))$. Strict held-out accuracy counts a score tie as an error, and the best and worst endpoint accuracies are its extrema over the endpoint set:

$$\mathrm{Acc}_T(\theta; C) = \frac{1}{|T|}\sum_{x \in T} \mathbf{1}\Big\{F_{\theta,C(x)}(x) > \max_{a \neq C(x)} F_{\theta,a}(x)\Big\},\ \ A_q^+ = \max_{\mathcal{S}_q} \mathrm{Acc}_T,\ \ A_q^- = \min_{\mathcal{S}_q} \mathrm{Acc}_T\,. \tag{5}$$

For a threshold $\tau \in (0,1)$ the endpoint set is in exactly one of three phases, each a quantifier over $\mathcal{S}_q$:

$$\textsf{all-good}: \ A_q^- \geq \tau, \qquad \textsf{mixed}: \ A_q^+ \geq \tau > A_q^-, \qquad \textsf{no-good}: \ A_q^+ < \tau. \tag{6}$$

Write $G_{q,\tau}^{\forall}, G_{q,\tau}^{\mathrm{mix}}, G_{q,\tau}^{\varnothing}$ for their indicators and $G_{q,\tau}^{\exists} = G_{q,\tau}^{\forall} + G_{q,\tau}^{\mathrm{mix}}$ for "some endpoint is good". Averaging over the mask, for $\star \in \{\forall, \mathrm{mix}, \varnothing, \exists\}$, gives the phase probabilities and the *phase profile*,

$$p_{q,\tau}^{\star}(C) = \mathbb{E}_{\Omega \sim \nu_q}\big[G_{q,\tau}^{\star}(C,\Omega)\big], \qquad \Pi_{q,\tau}(C) = \big(p_{q,\tau}^{\forall}(C),\, p_{q,\tau}^{\mathrm{mix}}(C),\, p_{q,\tau}^{\varnothing}(C)\big). \tag{7}$$

Proposition 3.3 (Phase decomposition)**.** *$\mathcal{S}_q(\mathcal{D})$ is non-empty for every $C \in \mathcal{C}_q$ and $\Omega \subseteq X_q$. For every $\tau \in (0,1)$ the three phases in* (6) *are mutually exclusive and exhaustive, so $\Pi_{q,\tau}(C)$ lies in the simplex $\Delta_2 = \{(a,b,c) \in [0,1]^3 : a+b+c = 1\}$.*

The threshold only decides where an instance falls; the extremes $A_q^- \leq A_q^+$ are properties of the set itself. A fixed instance is all-good for thresholds up to A_q^-, mixed between A_q^- and A_q^+, and no-good above, so it is mixed at *some* threshold exactly when different endpoints achieve different accuracies (Appendix B.1).

3.4 Invariance: the profile lives on colouring orbits

Permuting rows, columns, or symbols, or exchanging the two inputs, changes how a table is written and nothing about the learning problem. The formal name for this equivalence is colouring isomorphism (for Latin squares, isotopy together with transposition); the paper calls it relabelling.

Definition 3.4 (Colouring isomorphism)**.** $C \sim_q C'$ when there are $\rho, \kappa, \gamma \in S_q$ with $C'(u,v) = \gamma(C(\rho(u), \kappa(v)))$ for all u, v, or $C'(u,v) = \gamma(C(\kappa(v), \rho(u)))$ for all u, v. The quotients are $\widehat{\mathcal{Q}}_q = \mathcal{C}_q/\!\sim_q$ and $\mathcal{Q}_q = \mathcal{L}_q/\!\sim_q$.

Theorem 3.5 (Quotient invariance)**.** *Let $C \sim_q C'$. There are bijections of fixed-size masks and of endpoint sets, preserving the parameter norm and the mask law, under which corresponding endpoints have the same logits, the same truth–rival margins, and the same strict correctness at every cell. Consequently $p_{q,\tau}^{\forall}, p_{q,\tau}^{\mathrm{mix}}, p_{q,\tau}^{\varnothing}, p_{q,\tau}^{\exists}$ and $\Pi_{q,\tau}$ are well defined on $\widehat{\mathcal{Q}}_q$, hence on $\mathcal{Q}_q$, for every $\tau \in (0,1)$.*

The proof is short: relabelling the task amounts to permuting the entries of the two embedding tables and the rows of the head, and swapping the two tables under transposition, and each of these moves preserves the norm, the margin constraints, and cellwise correctness. The consequence is that the profile becomes a function on orbits, the *classification map*

$$\mathcal{Q}_q \longrightarrow \Delta_2, \qquad [C] \longmapsto \Pi_{q,\tau}([C]), \tag{8}$$

which is the object classified in the rest of the paper. One consequence deserves to be seen before the full proof (Appendix C.2): the support size and degree of the polynomial coordinate are not orbit invariants, so every relabelled representative $\gamma(\rho(u) + \kappa(v))$ of the addition orbit has the profile of C_q^{add}, however dense its polynomial.

Dense representatives of addition grok. Figure 2 is the invariance theorem seen from the data side. Ten random relabellings (ρ, κ, γ) of C_q^{add} at $q = 97$ give tables whose reduced polynomials have total degree 190 and 97% coefficient support, matching the random-matching Latin controls in both statistics, while their orbit is still that of C_q^{add}. All 30 runs memorise and then generalise, ending at median held-out accuracy 1.000; the 24 controls fit their training cells and end below 0.0014 held-out. Reusing a symbol already seen in the row or column of a cell is always wrong by the Latin property, so below-chance accuracy is possible. Within the orbit final accuracies cluster near one but the delays vary with the optimisation seed: a random-effects decomposition of the transition midpoint attributes about 93% of its variance to it (Appendix D.3). At this order, these runs support a stable final outcome with variable delay; they do not certify membership in $\mathcal{S}_q$ or measure its phase. The next section states the endpoint guarantees.

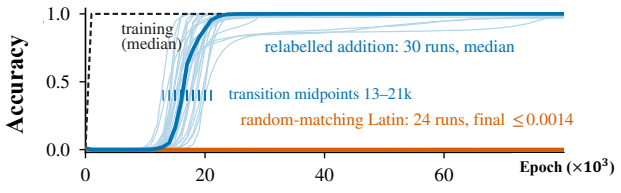


Figure 2: **Relabelled addition groks like addition.**

4 TWO POLES OF THE LATIN QUOTIENT

Informal Results. The objects of the previous section turn the question of the introduction into a map: each relabelling orbit of Latin colourings is sent to its phase profile, a point of the triangle whose corners are all-good, mixed, and no-good. Which orbits sit at the all-good corner? Which carry mixed mass? And where does the orbit of a uniformly random Latin colouring sit? Our results answer the first for every abelian group family and the third in its all-good coordinate, with the observation fraction and the success threshold held fixed while the order grows.

Informal Theorem (Two poles of the Latin quotient)**.** *Fix an observation fraction and a success threshold and let the order tend to infinity.*

(i) Abelian group families are all-good: *along any sequence of finite abelian groups of growing order, the worst minimum-norm endpoint accuracy converges to one in probability, so the profile converges to the all-good corner.*

(ii) Uniform Latin squares are unlearnable from the observations: *for a uniformly random Latin colouring, no learner whose prediction depends only on the observed cells reaches the success threshold with probability bounded away from zero. Consequently orbits with a non-negligible all-good component have vanishing mass under this law.*

Part (i) concerns every endpoint of each task in a structured family; part (ii) concerns every learner under a random task law. The latter controls the all-good coordinate, leaving open whether a typical colouring has some good endpoint.

4.1 THE MAIN RESULTS

We prove cyclic addition first and obtain every abelian group from it as a corollary, rather than proving the general statement directly: both proofs share one method, a carrier and a notch; the notch is insensitive to the colouring, and only the characters that build the carrier change from $\mathbb{Z}_q$ to a general abelian group. The cyclic case thus shows the whole argument in its most concrete form. Part (ii) needs two pieces of vocabulary: a task ensemble, since it concerns a random Latin square, and a learner, since it quantifies over everything that reads the observations. A *task ensemble* is a sequence $\mathsf{M} = (\mu_q)_q$ of distributions on $\mathcal{L}_q$; we use $\mathsf{M}_{\mathrm{add}} = (\delta_{C_q^{\mathrm{add}}})_q$ and $\mathsf{M}_{\mathrm{unif}} = (\mathrm{Unif}(\mathcal{L}_q))_q$. A *learner* A_q is a deterministic map from observed datasets to score tables on X_q, that is, to functions $X_q \to \mathbb{R}^q$ read through Acc_T as in (5) (a randomised learner is covered by conditioning on its coins). With $C \sim \mu_q$ and $\Omega \sim \nu_q$ independent,

$$P^{\star}_{q,\tau}(\mathsf{M}) = \mathbb{E}_{C\sim\mu_q}\big[p^{\star}_{q,\tau}(C)\big], \qquad P^{\mathsf{A}}_{q,\tau}(\mathsf{M}) = \mathbb{P}\big[\mathrm{Acc}_T(\mathsf{A}_q(\mathcal{D}(C,\Omega)); C) \geq \tau\big]. \tag{9}$$

A *selector* is a learner with $\mathsf{A}_q(\mathcal{D}) = F_\theta$ for some $\theta \in \mathcal{S}_q(\mathcal{D})$; since $\mathcal{S}_q(\mathcal{D})$ depends on the observations alone, any choice function on it is a learner, and one exists because the endpoint set is never empty.

Theorem 4.1 (Cyclic prototype and uniform obstruction)**.** *Fix $\alpha, \tau \in (0,1)$ and let $q \to \infty$.*

(i) (Addition is all-good.) *The worst endpoint accuracy converges to one in probability over the mask,*

$$A_q^-(C_q^{\mathrm{add}}, \Omega) \xrightarrow{\mathrm{P}} 1, \quad \textit{hence} \quad P^{\forall}_{q,\tau}(\mathsf{M}_{\mathrm{add}}) \to 1 \quad \textit{and} \quad \Pi_{q,\tau}([C_q^{\mathrm{add}}]) \to (1,0,0). \tag{10}$$

(ii) (Uniform Latin squares defeat every learner.) *Uniformly over all learners A_q,*

$$P^{\mathsf{A}}_{q,\tau}(\mathsf{M}_{\mathrm{unif}}) \;\leq\; (2e^2)^N\, q^{-\tau(1-\alpha)N} \;\longrightarrow\; 0, \tag{11}$$

hence $P^{\forall}_{q,\tau}(\mathsf{M}_{\mathrm{unif}}) \to 0$ and $\mathbb{P}_{C\sim\mathrm{Unif}(\mathcal{L}_q)}\big[p^{\forall}_{q,\tau}(C) \geq \varepsilon\big] \to 0$ for every $\varepsilon > 0$.

The two parts are proved in the next two subsections. The second part is pure counting and never looks at the model; the phase enters only through the observation that on an all-good instance every selector succeeds, so $P^{\forall}_{q,\tau}(\mathsf{M}_{\mathrm{unif}}) \leq P^{\mathsf{A}}_{q,\tau}(\mathsf{M}_{\mathrm{unif}})$, and Markov's inequality turns the vanishing expectation into the vanishing-mass statement. Nothing in either proof depends on the arithmetic of q: the theorem holds along every sequence of orders, with C_q^{add} the cyclic group table of order q. Written as $\exp\big(-N[\tau(1-\alpha)\log q - \log 2e^2]\big)$, the bound in (11) becomes informative once $q^{\tau(1-\alpha)} > 2e^2$; the statement about orbits is the push-forward of the uniform law on squares.

4.1.1 ADDITION: A CARRIER AND A NOTCH

The all-good pole rests on two lemmas. The first exhibits a cheap network that computes the whole addition table with a comfortable margin; the second says that, with high probability over the observed cells, no cheap network can agree with the observations and still be confidently wrong on a constant fraction of the table.

Lemma 4.2 (Addition carrier)**.** *There is an absolute constant C_0 such that for every sufficiently large q there is a full-grid score table $F^\star_q = A_{\mathrm{car}}\,\mathrm{ReLU}(Z_{\mathrm{car}}\xi_\cdot)$ of width $w_q = O(\log q)$ whose truth–rival gaps and product norm satisfy, for all $x \in X_q$ and $a \neq C^{\mathrm{add}}_q(x)$,*

$$1 \;\le\; F^\star_{q,C^{\mathrm{add}}_q(x)}(x) - F^\star_{q,a}(x) \;\le\; 9, \qquad J_{\mathrm{car}} := \|A_{\mathrm{car}}\|_F\,\|Z_{\mathrm{car}}\|_* \;\le\; C_0\, q\sqrt{\log q}. \tag{12}$$

The carrier is the Fourier-multiplication circuit of the interpretability literature, made explicit and costed. Roughly $\log q$ frequencies suffice for their averaged cosine to separate the correct sum from every wrong one by a constant gap; a small ReLU expansion of a piecewise-linear square approximates the required products of cosines and sines, and each of its hinges splits into a row part plus a column part, which is what the additive architecture needs. A constant number of hidden units per frequency gives product norm $O(q\sqrt{\log q})$ (Appendix C.3.1).

Lemma 4.3 (One-sided notch)**.** *Fix $\alpha, \beta \in (0,1)$ and let B_q be a deterministic sequence with $B_q^2 \log q / q^{5/2} \to 0$. With probability $1 - o(1)$ over $\Omega \sim \nu_q$ there is no table $D(x) = A\,\mathrm{ReLU}(Z\xi_x)$, of any finite width, with $\|A\|_F\|Z\|_* \le B_q$ that satisfies both*

(1) $D_{y(x)}(x) - D_a(x) \ge 0$ for every $x \in \Omega$ and $a \neq y(x)$, where $y(x) = C^{\mathrm{add}}_q(x)$; and

(2) at least βN cells x admit a rival $a(x) \neq y(x)$ with $D_{a(x)}(x) - D_{y(x)}(x) \ge 1$.

Both names are descriptive: the carrier carries the whole table at a known cost, and the notch is the region of the norm–error plane the lemma cuts out, cheap tables that fit the observations yet err confidently on a constant fraction of cells. The difficulty is that D may be chosen *after* the mask is seen, and the proof meets it in three steps. Before the mask is drawn it fixes a finite atlas $\mathcal{H}_q$ of suspect cell sets, such that the wrong cells of every cheap confidently-wrong table are covered, up to a vanishing fraction of N, by some member. The atlas is small: splitting Z at rank about $\sqrt{q}$, replacing both parts by a few sampled atoms, covering the sampled core by integer codes (Zhang, 2002; Ledent et al., 2021), and counting the sign patterns of the resulting polynomial gates (Caro & Datta, 2020) give $|\mathcal{H}_q| = \exp(o(N))$, a count that needs exactly the scaling assumed of B_q. And a fixed set of $\beta N/2$ cells meets a uniform mask in too few cells only with probability e^{-cN}, while a cheap table that is confidently wrong on an observed cell violates its margin there (Appendix C.3.2). Part (i) is therefore the competition

$$\mathbb{P}\big(\exists\ \text{harmful low-norm table}\big) \;\le\; |\mathcal{H}_q|\, e^{-cN} \;=\; \exp\big(\log|\mathcal{H}_q| - cN\big) \longrightarrow 0, \tag{13}$$

the carrier serving only to put every endpoint inside the budget B_q; width enters through the budget alone (Appendix E.1).

Proof of Theorem 4.1(i). Padded with zero hidden rows to width N, $(A_{\mathrm{car}}, Z_{\mathrm{car}})$ is feasible for the minimisation defining J_* on every addition mask, so $J_* \le J_{\mathrm{car}} \le C_0 q\sqrt{\log q}$, and by Lemma 3.2 each endpoint response F has a width-N reduced representation (A_F, Z_F) with $\|A_F\|_F\|Z_F\|_* = J_*$. Take $D = 9F - F^\star_q$ and $y(x) = C^{\mathrm{add}}_q(x)$. On an observed cell $F_y - F_a \ge 1$ against a carrier gap of at most 9, so $D_y - D_a \ge 0$; at a held-out non-strict error, including a tie, some rival has $F_a \ge F_y$ against a carrier gap of at least 1, so $D_a - D_y \ge 1$. Balancing both representations by homogeneity and stacking them, $\widetilde{A} = [\,3A_F \;\; -A_{\mathrm{car}}\,]$ and $\widetilde{Z} = [\,3Z_F;\; Z_{\mathrm{car}}\,]$ represent D with $\|\widetilde{A}\|_F^2 = 9J_* + J_{\mathrm{car}}$ and $\|\widetilde{Z}\|_* \le 3\sqrt{J_*} + \sqrt{J_{\mathrm{car}}} \le \sqrt{2}\sqrt{9J_* + J_{\mathrm{car}}}$, so

$$J(D) \le \|\widetilde{A}\|_F\|\widetilde{Z}\|_* \le \sqrt{2}\,(9J_* + J_{\mathrm{car}}) \le 10\sqrt{2}\,C_0\, q\sqrt{\log q} =: B_q. \tag{14}$$

Then $B_q^2 \log q / q^{5/2} = O((\log q)^2/\sqrt{q}) \to 0$, and the stack has width $N + w_q$, which Lemma 4.3 permits. An endpoint with $\mathrm{Acc}_T < \tau$ has more than $(1-\tau)|T| \ge (1-\tau)(1-\alpha)N$ non-strict errors, so with $\beta = \frac{1}{2}(1-\tau)(1-\alpha)$ its table D is excluded by the lemma. The exclusion event is a property of the mask, so it holds simultaneously for all endpoints: with probability $1 - o(1)$,

$A_q^-(C_q^{\mathrm{add}}, \Omega) \geq \tau$. This holds for every fixed $\tau < 1$, which is $A_q^- \xrightarrow{\mathrm{P}} 1$; the phase decomposition gives the profile limit. □

Corollary 4.4 (Abelian group families are all-good)**.** *Let $(G_q)_q$ be any sequence of finite abelian groups with $|G_q| = q \to \infty$, identify their elements with $[q]$, and let $C_{G_q}(u, v) = u + v$ be their group tables. For every fixed $\alpha, \tau \in (0, 1)$,*

$$A_q^-(C_{G_q}, \Omega) \xrightarrow{\mathrm{P}} 1, \qquad \Pi_{q,\tau}([C_{G_q}]) \to (1, 0, 0). \tag{15}$$

The conclusion holds for every relabelled representative.

Cyclic frequencies are the characters of $\mathbb{Z}_q$; replacing them by the characters of G_q preserves the carrier's width $O(\log q)$, gaps $[1, 9]$, and cost $O(q\sqrt{\log q})$, and the notch uses no group arithmetic, so the same budget proves the corollary. The general criterion and full character calculation are in Appendices E.3.1 and E.3.2.

4.1.2 Uniform Latin squares: a counting bound

Lemma 4.5 (Counting)**.** *Let $n_q = \lfloor \alpha N \rfloor$ and $r = \lfloor (1 - \tau)(N - n_q) \rfloor$. For every learner A_q,*

$$\textit{(a)} \quad P_{q,\tau}^{\mathsf{A}}(\mathsf{M}_{\mathrm{unif}}) \leq \frac{q^{n_q} \sum_{i=0}^{r} \binom{N-n_q}{i} (q-1)^i}{|\mathcal{L}_q|}, \qquad \textit{(b)} \quad |\mathcal{L}_q| \geq \frac{(q!)^{2q}}{q^{q^2}} \geq \left(\frac{q}{e^2}\right)^{q^2}. \tag{16}$$

The uniform pole needs no geometry. Part (a) fixes the mask, notes that the observed labels can take at most q^{n_q} values, and for each value counts the tables within Hamming distance r of what the learner predicts; part (b) is a lower bound on the number of Latin squares, obtained by extending a square row by row and applying the van der Waerden permanent bound to the admissible entries of each new row (Egorychev, 1981; Falikman, 1981; Gurvits, 2005); the bound is essentially sharp, since $|\mathcal{L}_q| = \left((1 + o(1))\, q/e^2\right)^{q^2}$ (van Lint & Wilson, 2001). The ratio of the two counts is the theorem.

Proof of Theorem 4.1(ii). Since $\sum_{i \leq r} \binom{N-n_q}{i} (q-1)^i \leq 2^N q^r$ and $N - n_q - r \geq \tau(N - n_q) \geq \tau(1 - \alpha)N$, the two parts of (16) give $P_{q,\tau}^{\mathsf{A}}(\mathsf{M}_{\mathrm{unif}}) \leq 2^N e^{2N} q^{-\tau(1-\alpha)N}$, uniformly in A_q. On an all-good instance every deterministic selector succeeds, so $P_{q,\tau}^{\forall}(\mathsf{M}_{\mathrm{unif}}) \leq P_{q,\tau}^{\mathsf{A}}(\mathsf{M}_{\mathrm{unif}})$ for any such selector. □

Remark 4.6 (Biases). With hidden bias b and output bias c added to (2) and to the penalised norm, both assertions of the theorem survive (Corollary B.4): a biased endpoint response has a bias-free analysis representation of product $O(q^{7/6}(\log q)^{1/3})$, so $B_q^2 \log q / q^{5/2} \to 0$ and (13) is unchanged.

Remark 4.7 (Vanishing observation fractions). Applying part (i) at $\alpha = 1/k$ and choosing orders beyond which the failure probability at accuracy $1 - 1/k$ is below $1/k$ yields $\alpha_q \downarrow 0$ along which $A_q^-(C_q^{\mathrm{add}}, \Omega) \xrightarrow{\mathrm{P}} 1$ still holds. This is existence without a rate; the rate the atlas calculation suggests, $\alpha_q \gg (\log q)^3/\sqrt{q}$, is recorded as a proof-guided scaling question in Appendix B.4, and the empirical thresholds of Table 3 fall from 0.33 at $q = 47$ to 0.20 at $q = 211$.

Remark 4.8 (What part (ii) leaves open). The existential probability $P_{q,\tau}^{\exists}(\mathsf{M}_{\mathrm{unif}}) = \mathbb{P}[\exists \theta \in \mathcal{S}_q(\mathcal{D}(C, \Omega)) : \mathrm{Acc}_T(\theta; C) \geq \tau]$ under $C \sim \mathrm{Unif}(\mathcal{L}_q)$, the mixed/no-good boundary of the triangle, has no known limit, nor is any sequence with non-vanishing mixed mass known; Appendix B.5 packages the programme. The non-abelian and non-group extensions of the carrier criterion also remain open.

5 Beyond the Poles: Structured Families at Composite Order

At $q = 64 = 2^6$ we compare the abelian all-good families of the previous section with non-abelian and non-group candidates, using product factors unavailable at prime order to measure partial label recovery. Runs follow Appendix D.1 with label-stratified masks (distinct from ν_q); they do not measure the phase of $\mathcal{S}_q$. C_r is the cyclic table, C_2^k the elementary abelian table, R_r a task-seeded Latin table built from random perfect matchings (Appendix D.3), and products are direct products.

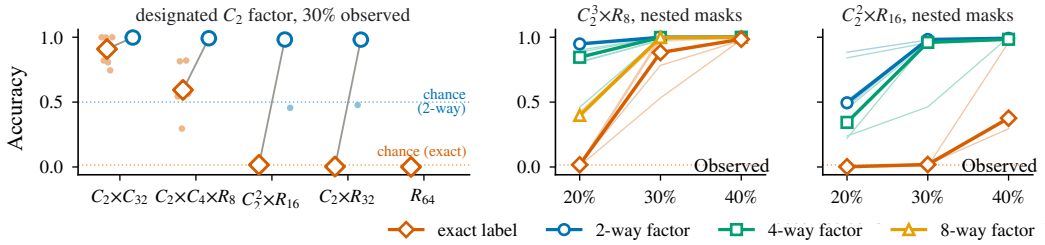


Figure 3: **Coarse factors are recovered before exact labels.** Colour encodes label resolution: exact label and the prespecified 2-, 4-, 8-way factor quotients. Left: held-out accuracy at nominal 30% coverage (runs small, medians outlined). Right: nested masks at 20%, 30%, 40%; coarse factors saturate one coverage level before the exact label.

Group laws beyond cyclic addition. Every group-table run in Table 1 transitions; C_2^6 completes exactly, and the other medians are 0.87–0.95. These finite-training outcomes need not match the asymptotic phase: 30% is at addition's training threshold at nearby orders (Table 3). $C_8 \times R_8$ yields one run above 0.99. The dihedral and non-group cases motivate further carriers (Appendix E.3).

Table 1: Final trained predictors at q=64, 30% coverage, six runs per family. "Transition" is the plateau-relative event of Appendix D.1; 0.99 is a reporting cutoff unrelated to τ.

	final outcome over 6 runs		
Latin family	trans.	med. acc.	≥ 0.99
C_2^6	6/6	1.000	6/6
C_{64}	6/6	0.949	3/6
$C_8\times C_8$	6/6	0.941	2/6
Dih_{64}	6/6	0.870	1/6
$C_8\times R_8$	4/6	0.231	1/6

Factor recovery with weak exact completion. The table $C_2 \times R_{32}$ carries a binary label component even when its 32-way residual is hard to infer. Runs are scored on exact labels and on *factor quotients*: projections of predicted and true labels to product factors, distinct from the orbit quotient $\mathcal{Q}_q$. In Figure 3, left, replacing the cyclic residual C_{32} by the random R_{32} lowers median exact recovery to chance or below, while binary accuracy remains above 0.98: one label component is recovered while finer components remain unresolved.

Coverage moves the recovered resolution. The nested-mask panels re-train the same task and optimisation seeds at increasing coverage. Coarse factors are recovered before exact labels, with the residual structure setting the required coverage: for $C_2^3 \times R_8$ the exact median rises $0.017 \to 0.882 \to 0.985$ while the binary factor is 0.948 already at 20%. This suggests a coordinate for the open region: the finest factor resolved at a given coverage (Appendix D.4). Checkpoint trajectories, alternate factor directions, and probe interventions (Appendix D.5) confirm that the factor accuracies reflect structured label resolution, not exact hits alone.

These observations motivate carriers beyond abelian groups and for coarse factors. For the all-good limit, we propose the following robust necessity conjecture.

Conjecture 5.1 (Robust margin necessity)**.** Let $C_q \in \mathcal{L}_q$ be deterministic with $q \to \infty$. If $A_q^-(C_q, \Omega) \xrightarrow{\mathrm{P}} 1$ for every fixed $\alpha \in (0, 1)$ under the corresponding uniform mask law, then C_q admits mask-independent carriers of width at most $N = q^2$, with truth–rival gaps at least 1 outside $o(N)$ cells and product norm $\|A_{\mathrm{car},q}\|_F \|Z_{\mathrm{car},q}\|_* = o(q^{3/2})$.

Appendix E.3.5 gives the precise formulation, motivation, and remaining proof gaps.

6 Conclusion

We study which task structures make the entire minimum-norm endpoint set generalise. Our carrier criterion guarantees this asymptotically for every finite abelian group family, whereas the uniform Latin ensemble defeats observation-based learners at fixed positive accuracy. Moreover, relabelling preserves endpoint generalisation despite changing polynomial complexity, making task structure central to this classification. Composite-order experiments further suggest recovery through coarse factors, without characterising the whole endpoint set. Together, these results connect grokking to the generalisation properties of its idealised endpoints, leaving carrier necessity and the mixed/no-good boundary as open questions.

AI USE STATEMENT

Generative AI tools were used to improve the language and clarity of the manuscript, and to surface existing results in the literature that shortened the proofs. The authors located the original sources and verified each result in the form used; the external proof ingredients are stated explicitly in Appendix C.4.

REFERENCES

Roman Abramov, Felix Steinbauer, and Gjergji Kasneci. Grokking in the wild: Data augmentation for real-world multi-hop reasoning with transformers. In *Int. Conf. Mach. Learn. (ICML)*, 2025. arXiv:2504.20752. 3

Hilal AlQuabeh, Velibor Bojkovic, Munachiso Nwadike, and Kentaro Inui. Mechanistic insights into grokking from the embedding layer. *arXiv preprint arXiv:2505.15624*, 2025. 3

Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, volume 35, pp. 21750–21764, 2022. 3

Alon Beck, Noam Levi, and Yohai Bar-Sinai. Grokking at the edge of linear separability. In *Int. Conf. Mach. Learn. (ICML)*, 2025. arXiv:2410.04489. 3

Matthias C. Caro and Ishaun Datta. Pseudo-dimension of quantum circuits. *Quantum Machine Intelligence*, 2(2):14, 2020. arXiv:2002.01490. 7, 25

Breno W. Carvalho, Artur S. d'Avila Garcez, Luís C. Lamb, and Emílio Vital Brazil. Grokking explained: A statistical phenomenon. *arXiv preprint arXiv:2502.01774*, 2025. 3

Kenzo Clauw, Sebastiano Stramaglia, and Daniele Marinazzo. Information-theoretic progress measures reveal grokking is an emergent phase transition. *arXiv preprint arXiv:2408.08944*, 2024. 3

Ben Cullen, Sergio Estan-Ruiz, Riya Danait, and Jiayi Li. A basin-selection perspective on grokking via singular learning theory. *arXiv preprint arXiv:2603.01192*, 2026. 3

Xander Davies, Lauro Langosco, and David Krueger. Unifying grokking and double descent. *arXiv preprint arXiv:2303.06173*, 2023. 3

G. P. Egorychev. The solution of van der Waerden's problem for permanents. *Advances in Mathematics*, 42(3):299–305, 1981. 8, 25

Ibrahim Talha Ersoy and Karoline Wiesner. Noise-driven escape from metastable phases explains grokking in deep neural networks. *arXiv preprint arXiv:2606.17120*, 2026. 3

Pavel Etingof, Oleg Golberg, Sebastian Hensel, Tiankai Liu, Alex Schwendner, Dmitry Vaintrob, and Elena Yudovina. Introduction to representation theory, 2011. URL `https://math.mit.edu/~etingof/replect.pdf`. Lecture notes, January 10. 26, 32

D. I. Falikman. Proof of the van der Waerden conjecture on the permanent of a doubly stochastic matrix. *Mathematical Notes of the Academy of Sciences of the USSR*, 29(6):475–479, 1981. 8, 25

Simin Fan, Razvan Pascanu, and Martin Jaggi. Deep grokking: Would deep neural networks generalize better? *arXiv preprint arXiv:2405.19454*, 2024. 3, 30

Hiroki Furuta, Gouki Minegishi, Yusuke Iwasawa, and Yutaka Matsuo. Interpreting grokked transformers in complex modular arithmetic. *arXiv preprint arXiv:2402.16726*, 2024. 3

Andrey Gromov. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*, 2023. 3

Zihan Gu, Ruoyu Chen, Hua Zhang, Yue Hu, and Xiaochun Cao. Beyond progress measures: Theoretical insights into the mechanism of grokking. *arXiv preprint arXiv:2504.03162*, 2025. 1, 3, 30, 31

Leonid Gurvits. A proof of hyperbolic van der Waerden conjecture: the right generalization is the ultimate simplification. *Electronic Colloquium on Computational Complexity (ECCC)*, 2005. Report TR05-103; arXiv:math/0504397. 8, 25

Tianyu He, Darshil Doshi, Aritra Das, and Andrey Gromov. Learning to grok: Emergence of in-context learning and skill composition in modular arithmetic tasks. *arXiv preprint arXiv:2406.02550*, 2024. 3

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. 18, 22, 25

Yufei Huang, Shengding Hu, Xu Han, Zhiyuan Liu, and Maosong Sun. Unified view of grokking, double descent and emergent abilities: A perspective from circuits competition. *arXiv preprint arXiv:2402.15175*, 2024. 3

Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. Deep networks always grok and here is why. In *Int. Conf. Mach. Learn. (ICML)*, 2024. 3

Hyeonbin Hwang and Yeachan Park. Intrinsic task symmetry drives generalization in algorithmic tasks. *arXiv preprint arXiv:2603.01968*, 2026. 3

Daniel Kunin, Allan Raventos, Clémentine Carla Juliette Dominé, Feng Chen, David Klindt, Andrew M Saxe, and Surya Ganguli. Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024. 3

Antoine Ledent, Waleed Mustafa, Yunwen Lei, and Marius Kloft. Norm-based generalisation bounds for deep multi-class convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8279–8287, 2021. arXiv:1905.12430. 7, 25

Jaerin Lee, Bong Gyun Kang, Kihoon Kim, and Kyoung Mu Lee. Grokfast: Accelerated grokking by amplifying slow gradients. *arXiv preprint arXiv:2405.20233*, 2024. 3

Ziyue Li, Chenrui Fan, and Tianyi Zhou. Grokking in LLM pretraining? monitor memorization-to-generalization without test. *arXiv preprint arXiv:2506.21551*, 2025. 3

Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, volume 35, pp. 34651–34663, 2022. 3

Róisín Luo, Christian Gagné, Jonas Ngnawé, Ihsan Ullah, and Karyn Morrissey. A stochastic–geometric theory of scaling laws in grokking. *arXiv preprint arXiv:2606.30388*, 2026. 3

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations (ICLR)*, 2020. 2, 3, 4

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon Shaolei Du, Jason D. Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *Int. Conf. Learn. Represent. (ICLR)*, 2024. 3

Neil Mallinar, Daniel Beaglehole, Libin Zhu, Adityanarayanan Radhakrishnan, Parthe Pandit, and Mikhail Belkin. Emergence in non-neural models: grokking modular arithmetic via average gradient outer product. *arXiv preprint arXiv:2407.20199*, 2024. 3

Shalima Binta Manir and Anamika Paul Rupa. A systematic empirical study of grokking: Depth, architecture, activation, and regularization. *arXiv preprint arXiv:2603.25009*, 2026. 3, 30

William Merrill, Nikolaos Tsilivis, and Aman Shukla. A tale of two circuits: Grokking as competition of sparse and dense subnetworks. *arXiv preprint arXiv:2303.11873*, 2023. 3

Jack Miller, Patrick Gleeson, Charles O'Neill, Thang Bui, and Noam Levi. Measuring sharpness in grokking. *arXiv preprint arXiv:2402.08946*, 2024. 3

Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica J Sutherland. Why do you grok? a theoretical analysis of grokking modular addition. *arXiv preprint arXiv:2407.12332*, 2024. 1, 3

Tiberiu Muşat. The geometry of grokking: Norm minimization on the zero-loss manifold. *arXiv preprint arXiv:2511.01938*, 2025. 3

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 1, 3

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, pp. 1376–1401, 2015. 2, 3, 4

Yeachan Park, Minseok Kim, and Yeoneung Kim. Acceleration of grokking in learning arithmetic operations via kolmogorov-arnold representation. *arXiv preprint arXiv:2405.16658*, 2024. 3

Gilles Pisier. Remarques sur un résultat non publié de B. Maurey. In *Séminaire d'Analyse Fonctionnelle (Polytechnique)*, pp. 1–12. École Polytechnique, 1981. Exposé no. V. 20, 25

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022. 1

Lucas Prieto, Melih Barsbey, Pedro A. M. Mediano, and Tolga Birdal. Grokking at the edge of numerical stability. In *Int. Conf. Learn. Represent. (ICLR)*, 2025. arXiv:2501.04697. 3

Jaisidh Singh, Diganta Misra, and Antonio Orvieto. Explaining grokking in transformers through the lens of inductive bias. *arXiv preprint arXiv:2602.06702*, 2026. 3

Yiding Song and Hanming Ye. Model capacity determines grokking through competing memorisation and generalisation speeds. *arXiv preprint arXiv:2605.09724*, 2026. 3

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018. 1, 3, 4

Hu Tan, Kuo Gai, and Shihua Zhang. Deciphering two training clocks in grokking via deep linear network theory with conditional ReLU reduction. *arXiv preprint arXiv:2606.05863*, 2026. 3

Yifan Tang, Qiquan Wang, Inés García-Redondo, and Anthea Monod. Topological signatures of grokking. *arXiv preprint arXiv:2605.06352*, 2026. 3

Vimal Thilak, Etai Littwin, Shuangfei Zhai, Omid Saremi, Roni Paiss, and Joshua M. Susskind. The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon. In *Has It Trained Yet? NeurIPS 2022 Workshop*, 2022. 3

Yuandong Tian. Provable scaling laws of feature emergence from learning dynamics of grokking. *arXiv preprint arXiv:2509.21519*, 2025. 3

Pascal Jr Tikeng Notsawo, Guillaume Dumas, and Guillaume Rabusseau. Grokking beyond the Euclidean norm of model parameters. In *Int. Conf. Mach. Learn. (ICML)*, 2025. arXiv:2506.05718. 3

Xuan Khanh Truong, Quynh Hoa Truong, Duc Trung Luu, and Thanh Duc Phan. The norm-separation delay law of grokking: A first-principles theory of delayed generalization. *arXiv preprint arXiv:2603.13331*, 2026. 3

J. H. van Lint and R. M. Wilson. *A Course in Combinatorics*. Cambridge University Press, 2nd edition, 2001. 8

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*, 2023. 3

Thomas Walker, Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. GrokAlign: Geometric characterisation and acceleration of grokking. *arXiv preprint arXiv:2506.12284*, 2025. 3

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization. *arXiv preprint arXiv:2405.15071*, 2024. 3

Hugh E. Warren. Lower bounds for approximation by nonlinear manifolds. *Transactions of the American Mathematical Society*, 133(1):167–178, 1968. 25

Mingyue Xu, Gal Vardi, and Itay Safran. To grok grokking: Provable grokking in ridge regression. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2026. arXiv:2601.19791. 3

Zhiwei Xu, Zhiyu Ni, Yixin Wang, and Wei Hu. Let me grok for you: Accelerating grokking via embedding transfer from a weaker model. In *Int. Conf. Learn. Represent. (ICLR)*, 2025. arXiv:2504.13292. 3

Alper Yıldırım. The geometric inductive bias of grokking: Bypassing phase transitions via architectural topology. *arXiv preprint arXiv:2603.05228*, 2026. 3

David Yunis, Kumar Kshitij Patel, Samuel Wheeler, Pedro Savarese, Gal Vardi, Karen Livescu, Michael Maire, and Matthew R Walter. Approaching deep learning through the spectral dynamics of weights. *arXiv preprint arXiv:2408.11804*, 2024. 3

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002. 7, 25

Zhangchen Zhou, Yaoyu Zhang, and Zhi-Qin John Xu. A rationale from frequency perspective for grokking in training neural network. *arXiv preprint arXiv:2405.17479*, 2024. 3

Xuekai Zhu, Yao Fu, Bowen Zhou, and Zhouhan Lin. Critical data size of language models from a grokking perspective. *arXiv preprint arXiv:2401.10463*, 2024. 3

A NOTATION

Table 2 lists the symbols used throughout the main text. Appendix-local symbols are defined where they enter.

Table 2: Notation used in the main text. Appendix-local symbols are defined where they enter.

Symbol	Meaning
$q, [q], S_q, X_q, N$	Task order, symbol set $\{0, \dots, q-1\}$, its symmetric group, cell grid $[q]^2$, and q^2.
$C, \mathcal{C}_q, \mathcal{L}_q, C_q^{\mathrm{add}}$	A colouring, all pairwise colourings, Latin colourings, and modular addition.
G_q, C_{G_q}	A finite abelian group of order q and its group table, after identifying its elements with $[q]$.
$\Omega, T, \nu_q, \alpha, n_q$	Observed cells, held-out cells, mask law, coverage, and mask size.
$\mathcal{D}(C, \Omega)$	Observed triples $(u, v, C(u, v))$.
$F_\theta, \theta, \mathcal{S}_q(\mathcal{D})$	Score map, network parameters, and minimum-norm endpoint set.
$\theta^{\mathrm{b}}, \mathcal{S}_q^{\mathrm{b}}, \Pi_{q,\tau}^{\mathrm{b}}$	Their L_2-regularised hidden/output-bias counterparts in Appendix B.2.
$\mathrm{Acc}_T, \tau, A_q^+, A_q^-$	Strict held-out accuracy, success threshold, and endpoint extrema.
$G_{q,\tau}^{\forall}, G_{q,\tau}^{\mathrm{mix}}, G_{q,\tau}^{\varnothing}, G_{q,\tau}^{\exists}$	All-good, mixed, no-good, and existential indicators.
$p_{q,\tau}^\star(C), \Pi_{q,\tau}(C)$	Threshold-indexed mask-averaged phase probability and phase profile.
$\sim_q, \widehat{\mathcal{Q}}_q, \mathcal{Q}_q$	Colouring isomorphism, ambient quotient space, and Latin quotient space.
ξ, Φ, Z	Cell indicator, stacked embedding matrix, and reduced hidden matrix $W\Phi$.
J, J_*, V_*	Reduced product functional, its constrained optimum, and the original norm optimum.
$F_q^\star, B_q$	Full-grid carrier score table and the product bound in the notch lemma.
$\mathsf{M}_{\mathrm{add}}, \mathsf{M}_{\mathrm{unif}}, P_{q,\tau}^\star(\mathsf{M})$	Addition and uniform-Latin task ensembles and their threshold-indexed probabilities.
$P_{q,\tau}^{\mathsf{A}}(\mathsf{M})$	Success probability of an observation-measurable learner A_q; a selector is a learner with values in $\mathcal{S}_q$.
C_r, C_2^k, R_r	Cyclic, elementary-abelian, and task-seeded sequential-Latin families used in Section 5.

B SUPPLEMENTARY THEORY

This appendix contains the exact finite-instance criterion behind the phases, the bias extension of Theorem 4.1, an unstructured comparison bound, and a formal statement of the classification programme that the two poles leave open. Proofs are in Appendix C.

B.1 AN EXACT CRITERION FOR A FINITE INSTANCE

Fix (q, C, Ω) and $\tau \in (0,1)$, and allow hidden width exactly N with zero padding. For $B \in \mathbb{R}^{N\times 2q}$ let $H(B)$ have x-th column $\mathrm{ReLU}(B\xi_x)$. On the nuclear sphere $\mathbb{S}_* = \{B : \|B\|_* = 1\}$ define

$$\rho(B) = \min_A \|A\|_F \quad \text{s.t.} \quad (e_{C(x)} - e_a)^\top A H(B)_x \geq 1 \quad (x \in \Omega,\ a \neq C(x)), \tag{17}$$

with $\rho(B) = +\infty$ if no A is feasible, and put $g = \min_{\mathbb{S}_*} \rho$ and $E_g = \{B \in \mathbb{S}_* : \rho(B) = g\}$. For $B \in E_g$ let A_B be the unique minimum-Frobenius solution of (17), $F_B = A_B H(B)$, $m_x(B) = F_{B,C(x)}(x) - \max_{a\neq C(x)} F_{B,a}(x)$ for $x \in T$, $s(B)$ the k-th largest held-out margin with $k = \lceil \tau|T| \rceil$, and

$$s_{-,\tau} = \min_{B\in E_g} s(B), \qquad s_{+,\tau} = \max_{B\in E_g} s(B). \tag{18}$$

Theorem B.1 (Exact finite-instance criterion). *E_g is non-empty and compact, $B \mapsto F_B$ is continuous on E_g, and the global responses of* (3) *are exactly $\{F_B : B \in E_g\}$. Moreover*

$$\textit{all-good} \iff s_{-,\tau} > 0, \qquad \textit{mixed} \iff s_{-,\tau} \leq 0 < s_{+,\tau}, \qquad \textit{no-good} \iff s_{+,\tau} \leq 0. \tag{19}$$

A tie $s = 0$ is on the bad side, matching (5). The criterion turns the phase of one instance into two extrema over a compact set.

Proposition B.2 (The mixed phase is a degeneracy question). *For fixed (C, Ω) the phases partition the threshold axis: all-good on $(0, A_q^-]$, mixed on $(A_q^-, A_q^+]$, no-good on $(A_q^+, 1)$. Hence*

$$\exists\, \tau \in (0,1) \textit{ with } (C,\Omega) \textit{ mixed} \iff \mathrm{Acc}_T(\cdot\,; C) \textit{ is non-constant on } \{F_B : B \in E_g\}. \tag{20}$$

If the held-out accuracy of the minimum-norm response is almost surely constant along a family, then $p^{\mathrm{mix}}_{q,\tau} \equiv 0$ on that family and its profile lies on the edge of Δ_2 joining $(1,0,0)$ to $(0,0,1)$ for every τ.

Two distinct responses of equal accuracy do not make an instance mixed, so (20) is strictly stronger than non-uniqueness of the minimum-norm response. A spread of accuracies across finite-training runs is not evidence of it either: such runs need not be global minimisers, and their spread is typically removable by changing the optimiser or regulariser. The natural source of the required degeneracy would be a nontrivial stabiliser of the observed data inside the isomorphism group of Definition 3.4, which by Theorem 3.5 carries one global response to another; we expect that stabiliser to be trivial with high probability under the uniform mask law, and responses that respect a symmetry produce score ties, which (5) counts as errors.

B.2 L_2-REGULARISED BIASES

Add a hidden bias $b \in \mathbb{R}^{h_q}$ and an output bias $c \in \mathbb{R}^q$, so that $\theta^{\mathrm{b}} = (E_U, E_V, W, A, b, c)$ and

$$F_{\theta^{\mathrm{b}}}(u,v) = A\ \mathrm{ReLU}\big(W(E_U(u) + E_V(v)) + b\big) + c, \qquad \|\theta^{\mathrm{b}}\|^2 = \|\theta\|^2 + \|b\|_2^2 + \|c\|_2^2, \tag{21}$$

and let $\mathcal{S}_q^{\mathrm{b}}(\mathcal{D})$ be the minimum-norm set under the constraints of (3). Superscript b marks the corresponding phase events, probabilities, and profile. With $Z = W\Phi$ and $\mathcal{F}_{\mathrm{b}}$ the set of (A, Z, b, c) satisfying the observed unit margins for $F(x) = A\,\mathrm{ReLU}(Z\xi_x + b) + c$, define

$$\Phi_{\mathrm{b}}(A, Z, b, c) = \|Z\|_* + \tfrac{1}{2}\big(\|A\|_F^2 + \|b\|_2^2 + \|c\|_2^2\big). \tag{22}$$

Lemma B.3 (Bias complexity bridge). *The biased problem is equivalent, in optimal value and in score responses, to $\min_{\mathcal{F}_{\mathrm{b}}} \Phi_{\mathrm{b}}$. There is an absolute constant C_{b} such that on every addition mask every biased minimum-norm response F has a finite-width bias-free representation $F(x) = \widehat{A}\,\mathrm{ReLU}(\widehat{Z}\xi_x)$ with $\|\widehat{A}\|_F \|\widehat{Z}\|_* \leq C_{\mathrm{b}} q^{7/6} (\log q)^{1/3}$.*

The representation is an analysis device whose width may exceed h_q by one unit, which Lemma 4.3 permits.

Corollary B.4 (Transfer to L_2-regularised biases)**.** *For the model* (21)*:*

(i) feasibility, phase decomposition, and quotient invariance hold, so $\Pi^{\mathrm{b}}_{q,\tau}([C])$ *is well defined on* $\widehat{\mathcal{Q}}_q$*;*

(ii) for fixed $\alpha, \tau \in (0,1)$*,* $\min_{\theta^{\mathrm{b}} \in \mathcal{S}^{\mathrm{b}}_q(\mathcal{D}(C_q^{\mathrm{add}},\Omega))} \mathrm{Acc}_T(\theta^{\mathrm{b}}; C_q^{\mathrm{add}}) \xrightarrow{\mathrm{P}} 1$*,* $P^{\forall,\mathrm{b}}_{q,\tau}(\mathsf{M}_{\mathrm{add}}) \to 1$*, and* $\Pi^{\mathrm{b}}_{q,\tau}([C_q^{\mathrm{add}}]) \to (1,0,0)$*;*

(iii) uniformly over deterministic biased selectors, $P^{\mathsf{A},\mathrm{b}}_{q,\tau}(\mathsf{M}_{\mathrm{unif}}) \le (2e^2)^N q^{-\tau(1-\alpha)N} \to 0$*; hence* $P^{\forall,\mathrm{b}}_{q,\tau}(\mathsf{M}_{\mathrm{unif}}) \to 0$ *and biased all-good colourings have zero uniform mass;*

(iv) on a finite instance, with $s_{-,\mathrm{b}}, s_{+,\mathrm{b}}$ *the extrema of the* $\lceil \tau|T| \rceil$*-th largest held-out margin over the biased reduced minimiser set, the three phases are characterised by the sign conditions of* (19)*; Theorem B.6 holds for the biased profiles.*

B.3 Unstructured colourings

Removing the Latin constraint gives an elementary independent-label bound for comparison with Lemma 4.5.

Proposition B.5 (Unstructured colourings)**.** *Let* $C \sim \mathrm{Unif}(\mathcal{C}_q)$*. For any learner, any* $\alpha \in (0,1)$*, and any* $\tau > 1/q$*,*

$$\mathbb{P}_{C\sim\mathrm{Unif}(\mathcal{C}_q),\,\Omega\sim\nu_q}\big[\mathrm{Acc}_T(\mathsf{A}_q(\mathcal{D}(C,\Omega)); C) \ge \tau\big] \le \exp\big(-2(\tau - 1/q)^2|T|\big). \tag{23}$$

Proof. Conditionally on $\mathcal{D}$, the labels on T are independent and uniform, so a strict prediction is correct with probability at most $1/q$ per cell, independently; Hoeffding's inequality gives the bound. □

B.4 A proof-guided observation-scale question

Tracking the discarded-cell parameter η through the atlas calculation of Appendix C.3.2 suggests that the notch argument tolerates

$$\alpha_q \gg \frac{(\log q)^3}{\sqrt{q}}, \qquad \text{equivalently} \qquad n_q \gg q^{3/2}(\log q)^3. \tag{24}$$

Turning this into a theorem requires a uniform accounting of the η-dependence in the covering and concentration bounds; the $O(q\,\mathrm{polylog}\,q)$ scale is a separate target.

B.5 A formal target for complete asymptotic classification

Fix $\tau \in (0,1)$. Let $K_{\infty,\tau}$ be the set of subsequential limits of $\Pi_{q,\tau}([C_{q_j}])$ over sequences of orders $q_j \to \infty$ and classes $[C_{q_j}] \in \mathcal{Q}_{q_j}$. Set $\mathsf{v} = (1,0,0)$ and $\mathsf{e} = \{(0,\lambda,1-\lambda) : 0 \le \lambda \le 1\}$, and define the uniform polarisation defect

$$U_{q,\tau} = \sup_{[C]\in\mathcal{Q}_q} p^{\forall}_{q,\tau}([C])\big(1 - p^{\forall}_{q,\tau}([C])\big). \tag{25}$$

Full edge realisation holds when for every $\lambda \in [0,1]$ there are $q_j \to \infty$ and classes with $p^{\forall}_{q_j,\tau}([C_{q_j}]) \to 0$ and $p^{\mathrm{mix}}_{q_j,\tau}([C_{q_j}]) \to \lambda$.

Theorem B.6 (Classification readiness)**.** *Theorem 4.1(i) gives* $\mathsf{v} \in K_{\infty,\tau}$*. Then* $K_{\infty,\tau} = \{\mathsf{v}\} \cup \mathsf{e}$ *holds if and only if both* $U_{q,\tau} \to 0$ *and full edge realisation holds. Under these two conditions a convergent profile sequence tends to* v *iff* $p^{\forall}_{q,\tau} \to 1$*, and to* $(0,\lambda,1-\lambda)$ *iff* $p^{\forall}_{q,\tau} \to 0$ *and* $p^{\mathrm{mix}}_{q,\tau} \to \lambda$*.*

The theorem packages the remaining programme as a polarisation estimate and a realisation construction. Realisation is the harder half: Proposition B.2 says it needs instances on which Acc_T is non-constant over the global responses, and none is known.

C PROOFS

This appendix gives the detailed proofs for Section 3, Theorem 4.1, and Appendix B; the abelian extension Corollary 4.4 is proved in Appendix E.3.2. Appendix C.1 covers the modelling section, Appendix C.2 the invariance theorem and the finite-instance criterion, Appendix C.3 the two poles of Theorem 4.1, Appendix C.3.5 the bias transfer, and Appendix C.4 states the external results in the forms used.

C.1 PROOFS FOR SECTION 3

Proof of Proposition 3.1. Assigning the value $C(u, v)$ to $U_u V_v$ is a bijection between label tables and edge colourings with colour set $[q]$, for every integer $q \geq 2$.

Now suppose q is a prime power and identify $[q]$ with $\mathbb{F}_q$. For $a \in \mathbb{F}_q$, define

$$\ell_a(t) = 1 - (t - a)^{q-1}. \tag{26}$$

The finite-field identity $z^{q-1} = 1$ for $z \neq 0$ gives $\ell_a(t) = \mathbf{1}\{t = a\}$ for $t \in \mathbb{F}_q$. Hence

$$P_C(X, Y) = \sum_{u,v \in \mathbb{F}_q} C(u, v)\, \ell_u(X) \ell_v(Y) \tag{27}$$

has degree at most $q - 1$ in each variable and agrees with C on $\mathbb{F}_q^2$. The vector space of such reduced polynomials and the vector space of functions $\mathbb{F}_q^2 \to \mathbb{F}_q$ both have dimension q^2. Since (27) makes the evaluation map surjective, it is bijective, proving uniqueness.

Suppose C is a Latin colouring. Fixing a left endpoint U_u, the colours $C(u, 0), \ldots, C(u, q - 1)$ are pairwise distinct because each symbol appears once in row u; fixing a right endpoint gives the same statement for columns. Hence edges sharing an endpoint receive distinct colours, which is properness.

Conversely, properness says the q edges at each left vertex, and at each right vertex, carry pairwise distinct colours. Since there are exactly q colours available, each colour occurs exactly once in every row and every column, which is the Latin condition.

Fix a colour a. The Latin condition gives each left vertex and each right vertex exactly one incident edge of colour a, so M_a is a perfect matching. The classes are pairwise disjoint and cover $E(K_{q,q})$ by construction. Conversely, if the edge set decomposes into q perfect matchings then each colour occurs exactly once at each vertex, which is properness and hence the Latin condition.

Under each bijection, the mask selects the same evaluations, cells, or edges. □

Proof of Proposition 3.3: non-emptiness. We exhibit a feasible point of finite norm; attainment then follows from closedness and coercivity. Let $e_0, \ldots, e_{2q}$ be the standard basis of $\mathbb{R}^{d_q}$, $d_q = 2q + 1$, and set

$$E_U(u) = e_u + \tfrac{1}{2} e_{2q}, \qquad E_V(v) = e_{q+v} + \tfrac{1}{2} e_{2q}, \tag{28}$$

so that $E_U(u) + E_V(v) = e_u + e_{q+v} + e_{2q}$. For each cell (i, j) place one hidden row

$$w_{ij} = e_i + e_{q+j} - \tfrac{3}{2} e_{2q}. \tag{29}$$

Its preactivation at cell (u, v) is $\mathbf{1}\{u = i\} + \mathbf{1}\{v = j\} - \frac{3}{2}$, which equals $\frac{1}{2}$ at (i, j) and is at most $-\frac{1}{2}$ elsewhere. Hence after ReLU the unit is active only at (i, j), with value $\frac{1}{2}$. Let this unit output with weight 2 to class $C(i, j)$ and 0 to every other class. The resulting network has score 1 at the true class of every cell and 0 at every other class, so the margin constraint of (3) is met with equality on the whole grid, in particular on Ω. There are $q^2 = h_q$ such units, so the width is respected and the norm is finite.

The feasible set of (3) is closed, being an intersection of preimages of closed sets under the continuous map $\theta \mapsto F_\theta$, and $\theta \mapsto \frac{1}{2}\|\theta\|^2$ is coercive. A coercive continuous function on a non-empty closed set attains its infimum, so $\mathcal{S}_q(\mathcal{D}) \neq \varnothing$. □

C.2 PROOFS OF INVARIANCE AND OF THE FINITE-INSTANCE CRITERION

Proof of Theorem 3.5. Consider first the direct case of Definition 3.4. Given Ω', set

$$\Omega = \{(\rho(u), \kappa(v)) : (u, v) \in \Omega'\}, \tag{30}$$

a bijection of the space of subsets of a fixed size, hence measure-preserving for ν_q. Given $\theta = (E_U, E_V, W, A)$ define

$$E'_U(u) = E_U(\rho(u)), \qquad E'_V(v) = E_V(\kappa(v)), \qquad W' = W, \tag{31}$$

and let A' have rows $A'_{\gamma(a),:} = A_{a,:}$ for $a \in [q]$. Then

$$F_{\theta',\gamma(a)}(u, v) = F_{\theta,a}(\rho(u), \kappa(v)). \tag{32}$$

Each transformation permutes the entries of a finite table or the rows of a matrix, so $\|\theta'\| = \|\theta\|$. By (32) the training constraints correspond term by term, and applying the same construction with the inverse permutations yields the inverse map; hence we have a norm-preserving bijection between feasible sets, which must map global minimisers onto global minimisers. At a test cell the true label is transported by the same γ, so strict argmax correctness is preserved cell by cell, and therefore so is Acc_T.

For the transposed case set $\Omega = \{(\kappa(v), \rho(u)) : (u, v) \in \Omega'\}$ and

$$E'_U(u) = E_V(\rho(u)), \qquad E'_V(v) = E_U(\kappa(v)), \tag{33}$$

with A' permuted by γ as before. The new preactivation at (u, v) equals the old one at $(\kappa(v), \rho(u))$. The two embedding tables enter (2) symmetrically, so swapping them is again norm-preserving, and the uniform mask law is invariant under these coordinate permutations. Hence all four events and their probabilities are unchanged. □

Proof of Proposition 3.3: decomposition. By the first part the set $\mathcal{S}_q(\mathcal{D})$ is non-empty. If it contains no good global minimiser then $A_q^+ < \tau$, which is no-good. Otherwise $A_q^+ \geq \tau$; if additionally there is no bad global minimiser then all of them are good, which is all-good, and if there is one then good and bad coexist, which is mixed. The three cases are mutually exclusive and exhaustive. Taking probabilities under ν_q gives $p^{\forall}_{q,\tau} + p^{\mathrm{mix}}_{q,\tau} + p^{\varnothing}_{q,\tau} = 1$. The threshold-axis partition of Proposition B.2 is the same statement read for fixed (C, Ω) and varying τ, and (20) follows from Theorem B.1, since the global responses are exactly $\{F_B : B \in E_g\}$. □

Proof of Lemma 3.2. If $\Omega = \varnothing$, the zero parameters are the unique minimiser and all stated quantities vanish. Assume henceforth that $\Omega \neq \varnothing$. For any factorisation $Z = W\Phi$ the nuclear-norm factorisation identity gives

$$\frac{\|W\|_F^2 + \|\Phi\|_F^2}{2} \;\geq\; \|W\Phi\|_* = \|Z\|_*, \tag{34}$$

with equality attained as follows: if $Z = U\Sigma V^\top$ is a compact SVD, take $W = U\Sigma^{1/2}$ and $\Phi = \Sigma^{1/2}V^\top$, so that $\|W\|_F^2 = \|\Phi\|_F^2 = \operatorname{tr}\Sigma = \|Z\|_*$. Since $\operatorname{rank} Z \leq 2q < d_q = 2q + 1$, this factorisation respects the embedding dimension, and padding with zero rows respects the hidden width.

Now fix a reduced representation (A, Z). Positive homogeneity of ReLU gives, for every $t > 0$,

$$A\,\mathrm{ReLU}(Z\xi_x) = (t^{-1}A)\,\mathrm{ReLU}\big((tZ)\xi_x\big), \tag{35}$$

so the score table, and hence feasibility, is unchanged under $(A, Z) \mapsto (t^{-1}A, tZ)$. Eliminating W and Φ via (34), the objective of (3) at the balanced point becomes

$$\phi(t) = t\,\|Z\|_* + \frac{\|A\|_F^2}{2t^2}. \tag{36}$$

Then $\phi'(t) = \|Z\|_* - \|A\|_F^2/t^3$, so the unique stationary point is $t^3 = \|A\|_F^2/\|Z\|_*$, it is a minimum since $\phi'' > 0$, and

$$\min_{t>0} \phi(t) = \tfrac{3}{2}\big(\|A\|_F\,\|Z\|_*\big)^{2/3}. \tag{37}$$

Minimising over all feasible reduced representations gives $V_* = \frac{3}{2}J_*^{2/3}$. At the optimal t, $\|A\|_F^2$, $\|W\|_F^2$, and $\|\Phi\|_F^2$ all equal $J_*^{2/3}$, which is the balanced factorisation. □

Proof of Theorem B.1. If $\Omega = \varnothing$, then $g = 0$, $E_g = \mathbb{S}_*$, and every $A_B = 0$; the only global response is zero and the claim follows directly. Assume $\Omega \neq \varnothing$ below. Positive homogeneity normalises any non-zero reduced representation (A, B) to $(\|B\|_* A,\ B/\|B\|_*)$ without changing either the scores or the product $\|A\|_F \|B\|_*$. Hence the minimal reduced product equals $\inf_{B \in \mathbb{S}_*} \rho(B)$.

The set $\mathbb{S}_*$ is a compact subset of a finite-dimensional space. The function ρ is lower semicontinuous: if $B_i \to B$ with $\liminf \rho(B_i) < \infty$, pass to a subsequence along which the fixed-list minimisers A_i are bounded and convergent; continuity of the constraints in (17) shows the limit is feasible for B, so $\rho(B) \leq \liminf \rho(B_i)$. The cell-selector construction of Proposition 3.3 provides at least one feasible list, so $g < \infty$ and the minimum is attained. As $\rho \geq g$ everywhere on $\mathbb{S}_*$, we may write $E_g = \{B : \rho(B) \leq g\}$, which is closed in a compact set, hence compact.

For fixed B the feasible A form a closed convex set and $\|A\|_F^2$ is strictly convex, so the minimiser A_B is unique. If $B_i \to B$ with all $B_i \in E_g$, then $\|A_{B_i}\|_F = g$ for all i; any convergent subsequence has a limit that is feasible at B with norm g, hence equals A_B by uniqueness. Therefore the whole sequence converges and $B \mapsto A_B$ is continuous; ReLU is continuous, so $B \mapsto F_B$ is continuous. A finite-order statistic of continuous functions is continuous, so the extrema $s_{-,\tau}, s_{+,\tau}$ in (18) are attained.

Any reduced global minimiser, once normalised, must have $B \in E_g$, and its output must be A_B: otherwise replacing it by the smaller-norm fixed-list output would strictly decrease the product, contradicting optimality. Conversely every $B \in E_g$ paired with A_B attains the minimal product and lifts, by the balancing step of Lemma 3.2, to a global minimiser of the original problem. Hence $\{F_B : B \in E_g\}$ is exactly the set of global responses.

Finally, a response has strict test accuracy at least τ iff at least $k = \lceil \tau |T| \rceil$ of its held-out margins are strictly positive, iff $s(B) > 0$. Therefore all global responses are good iff $\min_{E_g} s > 0$; good and bad coexist iff the minimum is non-positive while the maximum is positive; and no global is good iff the maximum is non-positive. An exact tie $s = 0$ counts as bad by the strict convention in (5). This is (19). □

C.3 Proofs for Section 4.1

C.3.1 The addition carrier

Proof of Lemma 4.2. We give the construction explicitly. Put

$$h = \tfrac{2}{3}, \qquad \epsilon = \frac{\sqrt{2}h^2}{4} = \frac{\sqrt{2}}{9}, \qquad \rho = \frac{1-\epsilon}{2}, \qquad c = \frac{2}{1-\epsilon}, \tag{38}$$

and

$$L = \left\lceil \frac{2(\log(q-1)+1)}{(\rho + 1/(q-1))^2} \right\rceil = O(\log q). \tag{39}$$

Frequency selection. We claim there are non-zero frequencies $k_1, \ldots, k_L \in \mathbb{Z}_q$ such that every $e \neq 0$ satisfies

$$\frac{1}{L} \sum_{\ell=1}^{L} \cos(2\pi k_\ell e/q) \ \leq \ \rho. \tag{40}$$

Argue probabilistically: for uniform non-zero k, the mean of $\cos(2\pi k e/q)$ is $-1/(q-1)$, since the full sum over all $k \in \mathbb{Z}_q$ vanishes. The summands lie in $[-1, 1]$, so Hoeffding's inequality (Hoeffding, 1963) bounds the probability that (40) fails at a fixed e by $\exp\{-L(\rho + 1/(q-1))^2/2\}$. A union bound over the $q-1$ values of e gives total probability below 1 by the choice of L, so the required frequencies exist.

An exact ReLU surrogate for multiplication. Let Q_h be the piecewise-linear interpolant of t^2 on the grid $\{-2, -\frac{4}{3}, -\frac{2}{3}, 0, \frac{2}{3}, \frac{4}{3}, 2\}$. It has the exact ReLU representation

$$Q_h(t) = h|t| + 2h \sum_{j=1}^{2} (|t| - jh)_+, \qquad |t| = (t)_+ + (-t)_+, \tag{41}$$

and satisfies $0 \le Q_h(t) - t^2 \le h^2/4$ for $|t| \le 2$. Hence

$$P(x,y) = \frac{Q_h(x+y) - Q_h(x-y)}{4} \tag{42}$$

satisfies, for $|x|, |y| \le 1$,

$$|P(x,y) - xy| \le h^2/16. \tag{43}$$

Scores. For each frequency k_ℓ, set $\phi_\ell(t) = 2\pi k_\ell t/q$ and use (43) inside the addition formulas:

$$c_\ell(u,v) = P(\cos\phi_\ell(u), \cos\phi_\ell(v)) - P(\sin\phi_\ell(u), \sin\phi_\ell(v)), \tag{44}$$

$$s_\ell(u,v) = P(\sin\phi_\ell(u), \cos\phi_\ell(v)) + P(\cos\phi_\ell(u), \sin\phi_\ell(v)), \tag{45}$$

so that $|c_\ell - \cos\phi_\ell(u+v)| \le h^2/8$ and $|s_\ell - \sin\phi_\ell(u+v)| \le h^2/8$. Define

$$F^\star_{q,a}(u,v) = \frac{c}{L} \sum_{\ell=1}^{L} \big(\cos\phi_\ell(a)\, c_\ell(u,v) + \sin\phi_\ell(a)\, s_\ell(u,v)\big). \tag{46}$$

Margins. The idealised version of (46), with c_ℓ, s_ℓ replaced by exact cosines and sines, scores c at $a = u + v$ and, by (40), at most $c\rho$ at any wrong class. Each realised score differs from its ideal by at most $c\sqrt{2}h^2/8$, so every truth–rival gap is at least $c(1 - \rho - \sqrt{2}h^2/4) = 1$, which is the lower bound in (12). For the upper bound, each ideal score is at most c in absolute value and each realised score deviates by at most $c\epsilon/2$; since $\epsilon = \sqrt{2}/9 < 1/6$ and $c = 2/(1-\epsilon) < 12/5$, every gap is at most

$$2c\Big(1 + \frac{\epsilon}{2}\Big) < \frac{24}{5} \cdot \frac{13}{12} = \frac{26}{5} < 9. \tag{47}$$

Additivity and width. Every hinge in (41) is the positive part of $r_u + s_v - t$. The constant t may be split, half into the row potential and half into the column potential, so each is a legitimate additive preactivation in the reduced coordinates $\xi_{(u,v)} = e_u + e_{q+v}$ of Section 3.2. Each frequency needs only a constant number of hidden rows, so the total width is $O(L) = O(\log q)$, which is below q^2 for large q.

Norms. Each preactivation row has Euclidean norm $O(\sqrt{q})$ over the $2q$ row/column potential coordinates, and each corresponding output column has Euclidean norm $O(\sqrt{q}/L)$. Hence

$$\|Z\|_* \le \sqrt{O(L)}\, \|Z\|_F = O(L\sqrt{q}), \qquad \|A\|_F = O(\sqrt{q/L}), \tag{48}$$

so, taking these constructed factors as $(A_{\mathrm{car}}, Z_{\mathrm{car}})$, $J_{\mathrm{car}} := \|A_{\mathrm{car}}\|_F \|Z_{\mathrm{car}}\|_* \le C_0 q\sqrt{\log q}$ after fixing an absolute C_0, which is (12). A compact SVD of Z_{car} finally yields admissible W and Φ as in Lemma 3.2. □

C.3.2 The one-sided notch lemma

Proof of Lemma 4.3. Throughout, $y(x) = C_q^{\mathrm{add}}(x)$.

Step 0: the design Gram matrix. The full addition design satisfies

$$\| \sum_{x \in X_q} \xi_x \xi_x^\top \|_{\mathrm{op}} = 2q, \tag{49}$$

because that Gram matrix is $\begin{pmatrix} qI & \mathbf{1}\mathbf{1}^\top \\ \mathbf{1}\mathbf{1}^\top & qI \end{pmatrix}$.

Step 1: low-rank split. Let $m = \lceil \sqrt{q} \rceil$ and split Z into its rank-m SVD core Z_m and the tail. Write the i-th column of A as a_i and the i-th row of Z_m as $z_i^{(m)}$, and put the non-zero terms in the form

$$A\,\mathrm{ReLU}(Z_m \xi_x) = \sum_i s_i u_i\, \mathrm{ReLU}(w_i^\top \xi_x), \tag{50}$$

with $\|u_i\|_2 = \|w_i\|_2 = 1$ and $s_i = \|a_i\|_2 \|z_i^{(m)}\|_2$, so that

$$\sum_i s_i \le \|A\|_F \|Z_m\|_F \le \|A\|_F \|Z\|_* \le B_q. \tag{51}$$

By (49) each normalised scalar gate has full-grid squared norm at most $2q$, so Maurey sampling in Hilbert space (Pisier, 1981) yields a proxy built from K_0 core atoms with full-grid squared vector error at most $2qB_q^2/K_0$.

The i-th tail atom is $a_i\big(\mathrm{ReLU}(z_i^\top \xi_x) - \mathrm{ReLU}((z_i^{(m)})^\top \xi_x)\big)$. Using that ReLU is 1-Lipschitz, Cauchy–Schwarz, and the singular-value tail bound $\|Z - Z_m\|_F^2 \le \|Z\|_*^2/(m+1)$ — which follows from $(m+1)\sigma_{m+1} \le \sum_{j \le m+1} \sigma_j \le \|Z\|_*$ and termwise summation — we get

$$\Big(\sum_i \|\text{tail atom}_i\|_{\ell_2(X_q;\mathbb{R}^q)}\Big)^2 \le 2q\|A\|_F^2 \|Z - Z_m\|_F^2 \le \frac{2qB_q^2}{m+1}, \tag{52}$$

and sampling K_1 tail atoms divides the right-hand side by K_1.

Step 2: sampling budgets. Set

$$\epsilon_q = \frac{B_q^2 \log q}{q^{5/2}}, \qquad r_q = \begin{cases} \min\{\log q,\ \epsilon_q^{-1/2}\}, & \epsilon_q > 0, \\ \log q, & \epsilon_q = 0, \end{cases} \tag{53}$$

and

$$K_0 = \max\{1, \lceil r_q B_q^2/q \rceil\}, \qquad K_1 = \max\{1, \lceil r_q B_q^2/(qm) \rceil\}. \tag{54}$$

Then $r_q \to \infty$ and $r_q \epsilon_q \to 0$ by the hypothesis $\epsilon_q \to 0$, and the combined full-grid squared vector error of the two proxies is $O(N/r_q) = o(N)$.

Write the core proxy as $\widetilde{D}_0 = B_0 H_0$, dividing each sampled core scalar gate by the fixed constant $\sqrt{2q}$ and multiplying that constant into the output column. The Maurey construction gives

$$\|B_0\|_F^2 \le a_q^2 := C_1 N/r_q, \tag{55}$$

and the normalised feature table satisfies $\sum_{x \in X_q} \|H_0(x)\|_2^2 \le K_0$, so after discarding at most ηN cells,

$$\|H_0(x)\|_2^2 \le b_q^2 := K_0/(\eta N). \tag{56}$$

The tail proxy remains explicit outside the covering theorem:

$$\widetilde{T}(x) = \sum_{j=1}^{K_1} v_j \big(\mathrm{ReLU}(r_j^\top \xi_x) - \mathrm{ReLU}((r_j')^\top \xi_x)\big), \tag{57}$$

which avoids hiding a mask-adaptive component inside a fixed-feature cover.

Step 3: a mask-independent cover code for the core. Apply Theorem C.1 with $p = r = 2$, using the block-diagonal lift over the q output columns in Theorem C.2; that is, represent one evaluation (x, c) by the feature

$$\zeta_{x,c} = e_c \otimes H_0(x) \in \mathbb{R}^{qK_0}. \tag{58}$$

On the retained cells $\|\zeta_{x,c}\|_2 \le b_q$, while the stacked output weight has Euclidean norm $\|B_0\|_F \le a_q$. Fix precision $\delta = 1/16$. Zhang's proof first fixes a deterministic threshold grid $\{\theta_s\}_{s=0}^{L_q^{\mathrm{grid}}} \subset [-a_q b_q - \delta/2,\ a_q b_q + \delta/2]$ with $L_q^{\mathrm{grid}} \le \lceil 4a_q b_q/\delta + 2 \rceil$; this grid is determined by q, η, B_q, r_q and is independent of the numerical feature table and mask.

Dispose first of the degenerate branch: if $\delta > a_q b_q$ then Cauchy–Schwarz bounds every core score coordinate by $a_q b_q < \delta$, so the single zero centre is already a valid cover; add it to the atlas as a deterministic code. Assume henceforth $\delta \le a_q b_q$.

At $p = 2$ the duality map in Zhang's argument is linear, so each cover centre is determined by a finite non-negative integer multiplicity code $\kappa = (\alpha_{x,c,s}, \beta_{x,c,s})$ of total mass at most

$$M_q := \lceil 36 a_q^2 b_q^2 \delta^{-2} \rceil, \tag{59}$$

defined on the ambient address set $X_q \times [q] \times \{0, \dots, L_q^{\mathrm{grid}}\} \times \{+, -\}$ after zero-filling unused addresses. Zhang's augmented separating vector is

$$\begin{aligned} y_\kappa(H_0) &= 2 \sum_{x,c,s} (\alpha_{x,c,s} - \beta_{x,c,s})\, \zeta_{x,c}, \\ a_q z_\kappa &= \frac{2}{a_q} \sum_{x,c,s} (-\alpha_{x,c,s} + \beta_{x,c,s})\, \theta_s, \end{aligned} \tag{60}$$

where z_κ depends only on the code and the deterministic grid. Zhang's separation argument shows every code that actually produces a cover centre has $z_\kappa > 0$; in the enumeration we simply discard codes with $z_\kappa \le 0$, so no division by zero or sign ambiguity arises. The effective centre is $\widehat{w}_\kappa(H_0) = y_\kappa(H_0)/z_\kappa$. If $a_q = 0$ the core output vanishes identically and the zero centre suffices, so the formula is only used when $a_q > 0$; a sampled gate that vanishes on the whole grid contributes a zero feature and requires no division by a feature norm. By Theorems C.1 and C.2, every realised $B_0 H_0$ is approximated within δ at every output coordinate of every retained cell by some effective code centre.

The number of codes is likewise fixed before the mask is drawn. The address count is at most $L_q^{\mathrm{addr}} = 2Nq(L_q^{\mathrm{grid}} + 1)$, and the number of non-negative integer codes of total mass at most M_q is at most $(L_q^{\mathrm{addr}} + 1)^{M_q}$. From (54)–(56),

$$a_q^2 b_q^2 = O_\eta\Big(\frac{B_q^2}{q} + \frac{1}{r_q}\Big), \tag{61}$$

and $a_q b_q$ is at most a fixed polynomial in q, so

$$\log \#\{\kappa\} = O_\eta\Big(\Big(1 + \frac{B_q^2}{q} + \frac{1}{r_q}\Big) \log q\Big) = o(N). \tag{62}$$

The additive constant 1 is the at most one extra atom created by the ceiling in (59); it contributes $O(\log q) = o(N)$, and in the non-degenerate branch $a_q^2 b_q^2 \ge \delta^2$, so it is absorbed by the main term.

Step 4: sign-pattern counting. Write the common rank-m row space of the core as $R \in \mathbb{R}^{m \times 2q}$ and the j-th sampled core normal as $R^\top t_j$, giving $2qm + mK_0$ continuous core parameters; the tail atoms of (57) contribute at most $5qK_1$ parameters through $v_j \in \mathbb{R}^q$ and $r_j, r_j' \in \mathbb{R}^{2q}$. Hence

$$D_q^{\mathrm{par}} \le 2qm + mK_0 + 5qK_1 = O\Big(q^{3/2} + \frac{r_q B_q^2}{\sqrt{q}} + q\Big), \tag{63}$$

and therefore

$$\frac{D_q^{\mathrm{par}} \log q}{N} = O\Big(\frac{\log q}{\sqrt{q}} + r_q \epsilon_q\Big) = o(1). \tag{64}$$

For each x and j, the core preactivation $t_j^\top R \xi_x$ is a degree-2 polynomial in the continuous parameters, and the tail preactivations $r_j^\top \xi_x$, $(r_j')^\top \xi_x$ are degree-1. By Theorem C.3, the ternary sign patterns of these $(K_0 + 2K_1)N$ polynomials give at most $\exp(O(D_q^{\mathrm{par}} \log q))$; because zeros are counted, all exact zero gates are retained.

On a fixed gate stratum each $H_{0,j}(x)$ is a polynomial of degree at most 2, so by (60) the c-th score of an effective code centre at cell x has the explicit form

$$\widehat{F}_{\kappa,c}(x) = \frac{2}{z_\kappa} \sum_{s,c',\ell} \sigma_{\kappa,s,c',\ell}\, H_{0,\ell}(s) H_{0,\ell}(x), \tag{65}$$

with integer coefficients $\sigma_{\kappa,s,c',\ell}$ determined entirely by the code and free of continuous parameters. Thus (65) is a polynomial of degree at most 4 divided by the positive code-only constant z_κ. The tail score (57) is a polynomial of degree at most 2 on a fixed tail-gate stratum. Clearing the positive denominator, each truth–rival half-gap comparison

$$\widehat{F}_{\kappa,a}(x) + \widetilde{T}_a(x) - \widehat{F}_{\kappa,y(x)}(x) - \widetilde{T}_{y(x)}(x) - \tfrac{1}{2} \tag{66}$$

agrees in sign with a polynomial of degree at most 4, and there are only $N(q-1)$ such comparisons. Applying Theorem C.3 again — or the direct 3^M bound when the number of polynomials is below

the number of parameters — bounds the number of threshold sets per pair (code, gate stratum) by $\exp(O(D_q^{\mathrm{par}} \log q))$.

Combining (62) and (64), the deterministic atlas $\mathcal{H}_q$ of all threshold sets arising from gates, zero gates, score ties and half-gap comparisons satisfies

$$\log|\mathcal{H}_q| \le \log \#\{\kappa\} + O(D_q^{\mathrm{par}} \log q) = o(N). \tag{67}$$

Its address set, threshold grid, codes and polynomial list are all fixed before Ω is drawn, whereas the endpoint, the SVD row space, the sampled atoms, the tail atoms and the effective code may all be chosen after seeing Ω; the enumeration (67) covers them simultaneously.

Step 5: concentration. Let $\mathcal{X}_{\mathrm{exc}}$ contain the high-feature cells discarded in (56) together with the cells where either Maurey approximation exceeds the fixed vector threshold $1/16$. By Markov's inequality and the total squared errors of Step 2, choosing $\eta > 0$ small and then letting $q \to \infty$,

$$|\mathcal{X}_{\mathrm{exc}}| \le (\eta + o(1))N. \tag{68}$$

Off $\mathcal{X}_{\mathrm{exc}}$ the core Maurey error and the tail Maurey error each have vector norm below $1/16$ and the Zhang centre has entrywise error below $1/16$, so the final proxy $\bar{D}$ differs from D by less than $3/16$ in each coordinate and by less than $3/8$ in each pair gap. Define the atlas threshold set

$$H_{\bar{D}} = \Big\{x : \max_{a \ne y(x)} \big(\bar{D}_a(x) - \bar{D}_y(x)\big) \ge \tfrac{1}{2}\Big\}. \tag{69}$$

Hypothesis (1) gives $|H_{\bar{D}} \cap \Omega| \le |\mathcal{X}_{\mathrm{exc}}|$, while the βN reverse-unit-gap cells of hypothesis (2) give $|H_{\bar{D}}| \ge \beta N - |\mathcal{X}_{\mathrm{exc}}| \ge \beta N/2$ for large q.

Fix any $H \subseteq X_q$ with $|H| \ge \beta N/2$. Then $|H \cap \Omega|$ is hypergeometric with mean at least $\alpha\beta N/2 + O(1)$. Choosing $0 < \eta < \alpha\beta/16$, the binomial-coefficient formula gives

$$\mathbb{P}\big(|H \cap \Omega| \le 2\eta N\big) = \sum_{j \le 2\eta N} \frac{\binom{|H|}{j}\binom{N-|H|}{n_q - j}}{\binom{N}{n_q}} \le e^{-c_{\alpha,\beta,\eta} N} \tag{70}$$

for some $c_{\alpha,\beta,\eta} > 0$; the same bound follows from sampling-without-replacement Chernoff estimates (Hoeffding, 1963). Taking a union bound over the $\exp(o(N))$ sets H permitted by (67) leaves total probability $o(1)$. Hence with probability $1 - o(1)$ no such D exists. Exact zero gates and exact ties are preserved throughout by the ternary sign counting and by the weak threshold $\frac{1}{2}$. □

C.3.3 Assembly of part (i)

The proof of Theorem 4.1(i) is given in full in Section 4.1.1. For the bias transfer of Appendix C.3.5 we record the two gap inequalities in the form used there. With $D = 9F - F_q^\star$, $y = C_q^{\mathrm{add}}(x)$, and $\gamma_{x,a} = F_{q,y}^\star(x) - F_{q,a}^\star(x) \in [1, 9]$ the carrier gap,

$$D_y - D_a = 9(F_y - F_a) - \gamma_{x,a} \ge 0 \qquad \text{on observed cells, since } F_y - F_a \ge 1, \tag{71}$$

and at every held-out non-strict error, including a tie, some $a \ne y$ has $F_a \ge F_y$, whence

$$D_a - D_y = 9(F_a - F_y) + \gamma_{x,a} \ge 1. \tag{72}$$

The norm budget is B_q of (14).

C.3.4 The uniform Latin square obstruction

Proof of Lemma 4.5(a). Since C is uniform on $\mathcal{L}_q$ and independent of Ω, $P_{q,\tau}^{\mathsf{A}}(\mathsf{M}_{\mathrm{unif}}) = \mathbb{E}_\Omega\big[|\{|C \in \mathcal{L}_q : \mathrm{Acc}_T(\mathsf{A}_q(\mathcal{D}(C,\Omega)); C) \ge \tau\}/|\mathcal{L}_q|\big]$, so it suffices to bound the count for a fixed mask Ω. The observed labels take at most q^{n_q} values. Fix one observed dataset $\mathcal{D}$. On each test cell the learner has at most one strict winner; if there is no strict winner, assigning an arbitrary label only enlarges the candidate set counted below.

Success means the true Latin table differs from the winner table on at most r test cells. Choosing those positions and their replacement labels, the number of full arrays within that Hamming ball is at most

$$\sum_{i=0}^{r} \binom{N - n_q}{i} (q-1)^i, \tag{73}$$

and restricting to Latin arrays can only decrease the count. Summing over all $\mathcal{D}$ and averaging over fixed-size masks gives (16). □

Proof of Lemma 4.5(b). By the permanent bound in Theorem C.4, if B is a $q \times q$ non-negative doubly stochastic matrix then $\mathrm{per}(B) \geq q!/q^q$.

Build the Latin square row by row. When k rows have been placed, the symbol–column incidence matrix of admissible entries for the next row is $(q-k)$-regular; dividing by $q-k$ makes it doubly stochastic, so its permanent is at least

$$(q-k)^q \, q!/q^q. \tag{74}$$

The first row admits $q!$ choices, so

$$|\mathcal{L}_q| \;\geq\; q! \prod_{k=1}^{q-1} \frac{(q-k)^q q!}{q^q} = \frac{(q!)^{2q}}{q^{q^2}}. \tag{75}$$

Finally $\log(q!) = \sum_{j=1}^{q} \log j \geq \int_1^q \log x \, dx = q \log q - q + 1$ gives $q! \geq (q/e)^q$, and substituting yields $(q/e^2)^{q^2}$. □

C.3.5 L_2-REGULARISED BIASES

Proof of Lemma B.3. For fixed $Z = W\Phi$, the nuclear-norm factorisation identity used in (34) gives

$$\inf_{W\Phi=Z} \frac{1}{2}\big(\|W\|_F^2 + \|\Phi\|_F^2\big) = \|Z\|_*. \tag{76}$$

Every admissible Z has at most $2q$ columns, while $d_q = 2q+1$, so an SVD factorisation attaining (76) respects the embedding dimension; zero padding respects the hidden width. Keeping A, b, c fixed therefore establishes both directions of the equivalence with $\min_{\mathcal{F}_\mathrm{b}} \Phi_\mathrm{b}$. Its feasible set is non-empty because the bias-free cell-selector construction of Proposition 3.3 is feasible with $b = c = 0$. Closedness and coercivity give attainment.

Let $F_q^\star$ be the addition carrier and let J_{car} denote the cost of the constructed representation in Lemma 4.2. Balancing its two reduced factors and taking $b = c = 0$ provides a feasible competitor for every addition mask. Hence the biased optimum M_q satisfies

$$M_q := \min_{\mathcal{F}_\mathrm{b}} \Phi_\mathrm{b} \leq \frac{3}{2} J_{\mathrm{car}}^{2/3} \leq C_1 q^{2/3} (\log q)^{1/3} \tag{77}$$

for an absolute constant C_1. Consequently, every reduced global minimiser obeys

$$\|Z\|_* \leq M_q, \qquad \|A\|_F, \|b\|_2, \|c\|_2 \leq \sqrt{2M_q}. \tag{78}$$

The hidden bias can be absorbed into the two endpoint blocks. Put

$$\overline{Z} = Z + \tfrac{1}{2} b \mathbf{1}_{2q}^\top. \tag{79}$$

Since ξ_x has exactly two unit entries, $\overline{Z}\xi_x = Z\xi_x + b$. The added rank-one matrix has nuclear norm $\sqrt{q/2}\,\|b\|_2$, and thus

$$\|\overline{Z}\|_* \leq M_q + \sqrt{qM_q}. \tag{80}$$

It follows that $F_0(x) = A\,\mathrm{ReLU}(\overline{Z}\xi_x)$ has a bias-free representation of cost at most

$$J(F_0) \leq \sqrt{2M_q}\big(M_q + \sqrt{qM_q}\big). \tag{81}$$

The constant table $F_c(x) = c$ has a one-hidden-unit representation: take $z_c = \frac{1}{2}\mathbf{1}_{2q}^\top$, for which $\mathrm{ReLU}(z_c\xi_x) = 1$, and use c as the output column. By (78),

$$J(F_c) \leq \sqrt{qM_q}. \tag{82}$$

If F_1, F_2 have respective product costs J_1, J_2, balance their factors, concatenate the output matrices, and stack the hidden matrices. The triangle and Cauchy–Schwarz inequalities give

$$J(F_1 + F_2) \leq (\sqrt{J_1} + \sqrt{J_2})^2 \leq 2(J_1 + J_2). \tag{83}$$

Applying (83) to $F = F_0 + F_c$ and using (77) yields

$$J(F) \le 2\left[\sqrt{2M_q}\big(M_q + \sqrt{qM_q}\big) + \sqrt{qM_q}\right] \tag{84}$$

$$= O\Big(M_q^{3/2} + \sqrt{q}\, M_q + \sqrt{qM_q}\Big) \tag{85}$$

$$= O\Big(q^{7/6}(\log q)^{1/3}\Big), \tag{86}$$

which proves the bound of Lemma B.3. □

Proof of Corollary B.4. The bias-free feasible set embeds into the biased one by taking $b = c = 0$, so feasibility persists; attainment again follows from closedness and coercivity. For a direct colouring isomorphism, use the parameter map in the proof of Theorem 3.5, keep b fixed, and permute the coordinates of c by the same symbol permutation used on the rows of A. Under transposition, also exchange the two embedding tables. These maps preserve (21), the constraints, and cellwise strict correctness, and their inverses have the same form. This proves the biased phase and quotient claims.

For addition, let F be any biased minimum-norm response. By Lemma B.3, it has an analysis representation with

$$J(F) = O\Big(q^{7/6}(\log q)^{1/3}\Big). \tag{87}$$

Set $D = 9F - F_q^\star$. The observed non-negative gaps in (71) and the unit reverse gaps at every held-out non-strict error in (72) depend only on the score table, so they remain valid. Applying (83) gives

$$J(D) \le 2\big(9J(F) + J(F_q^\star)\big) = O\Big(q^{7/6}(\log q)^{1/3}\Big), \tag{88}$$

and therefore

$$\frac{J(D)^2 \log q}{q^{5/2}} = O\Big(q^{-1/6}(\log q)^{5/3}\Big) \to 0. \tag{89}$$

For any fixed $\tau < 1$, an endpoint below accuracy τ supplies more than $(1-\tau)(1-\alpha)N$ reverse-gap cells. Taking $\beta = \frac{1}{2}(1-\tau)(1-\alpha)$, the one-sided notch lemma excludes all such biased endpoint responses simultaneously with probability $1 - o(1)$. Applying this statement at each fixed $\tau < 1$ proves the first convergence in Corollary B.4(ii); the all-good probability and profile follow from the phase decomposition.

The proof of Lemma 4.5 uses only observation measurability of the predicted table. It therefore applies verbatim to every biased minimum-norm selector and gives Corollary B.4(iii). On a biased all-good instance every such selector succeeds, which proves $P_{q,\tau}^{\forall,\mathrm{b}}(\mathsf{M}_{\mathrm{unif}}) \to 0$; Markov's inequality then gives the biased zero-mass statement.

For the finite-instance claim, let

$$g_{\mathrm{b}} = \min_{\mathcal{F}_{\mathrm{b}}} \Phi_{\mathrm{b}}, \qquad E_{g,\mathrm{b}} = \{(A, Z, b, c) \in \mathcal{F}_{\mathrm{b}} : \Phi_{\mathrm{b}}(A, Z, b, c) = g_{\mathrm{b}}\}. \tag{90}$$

This set is non-empty and compact. By (76), its score responses are exactly the original biased global responses. The held-out margins and their finite order statistics are continuous on $E_{g,\mathrm{b}}$. Taking the minimum and maximum of the k-th largest margin, where $k = \lceil \tau|T| \rceil$, and repeating the final step of Theorem B.1 gives the same three sign conditions. Finally, the proof of Theorem B.6 uses only the simplex identity, the addition vertex, and the definitions of the polarisation defect and edge realisation. The biased versions established above supply these ingredients, completing the transfer. □

C.3.6 Classification readiness

Proof of Theorem B.6. Assume first that $U_{q,\tau} \to 0$. Take any convergent profile sequence and let t be the limit of its first coordinate. By (25), $t(1-t) = 0$, so $t \in \{0, 1\}$. If $t = 1$ the simplex identity of Proposition 3.3 forces the limit to be v; if $t = 0$ the limit lies in e. Hence $K_{\infty,\tau} \subseteq \{\mathsf{v}\} \cup \mathsf{e}$. Theorem 4.1(i) gives $\mathsf{v} \in K_{\infty,\tau}$ and full edge realisation gives $\mathsf{e} \subseteq K_{\infty,\tau}$, so $K_{\infty,\tau} = \{\mathsf{v}\} \cup \mathsf{e}$ holds.

Conversely suppose $K_{\infty,\tau} = \{\mathsf{v}\} \cup \mathsf{e}$ holds and $\limsup_{q\to\infty} U_{q,\tau} > 0$. Then there are $\delta > 0$, a sequence of orders $q_j \to \infty$ and classes $[C_{q_j}]$ with

$$p_{q_j,\tau}^{\forall}([C_{q_j}])\big(1 - p_{q_j,\tau}^{\forall}([C_{q_j}])\big) \ge \delta. \tag{91}$$

The probability simplex is compact, so we may extract a further subsequence along which the profile converges. Its first coordinate lies strictly inside $(0, 1)$, so the limit is neither v nor a point of e, contradicting $K_{\infty,\tau} = \{\mathsf{v}\} \cup \mathsf{e}$. Therefore $U_{q,\tau} \to 0$. Moreover $K_{\infty,\tau} = \{\mathsf{v}\} \cup \mathsf{e}$ says every point of e has a realising sequence of classes by the definition of $K_{\infty,\tau}$, which is exactly full edge realisation.

The final coordinate characterisation is immediate from $p^{\forall}_{q,\tau} + p^{\mathrm{mix}}_{q,\tau} + p^{\varnothing}_{q,\tau} = 1$. □

C.4 External results used

We state the covering/counting results, permanent bound, sampling estimates, and character identities in the forms used in the proofs.

Theorem C.1 (Zhang, 2002, Theorem 4)**.** *Let sample vectors satisfy $\|x_i\|_p \le b$ and linear weights satisfy $\|w\|_r \le a$, where $2 \le p < \infty$ and $1/p + 1/r = 1$. Then for every $\varepsilon > 0$ the empirical ℓ_∞ covering number of the linear function class on these n samples satisfies*

$$\log_2 N_\infty \le 36(p-1)a^2b^2\varepsilon^{-2} \log_2\big(2\lceil 4ab/\varepsilon + 2\rceil n + 1\big). \tag{92}$$

Theorem C.2 (Ledent et al., 2021, Proposition 5)**.** *Let $X \in \mathbb{R}^{n\times U\times d}$ with $\|X_{i,u,:}\|_2 \le b$ for every patch. For any reference M, the matrix class $\{XA : \|A - M\|_F \le a\}$ admits an entrywise ℓ_∞ cover over all sample, patch and output-channel coordinates with*

$$\log N \le 36a^2b^2\varepsilon^{-2} \log_2\big(((8ab/\varepsilon) + 7)\, mnU\big), \tag{93}$$

where m is the number of output channels.

Theorem C.2 is the block-diagonal lift of Theorem C.1 to the index triple (sample, patch, output channel), which is exactly the form we need in Step 3 of the proof of Lemma 4.3.

Theorem C.3 (Caro & Datta, 2020, Corollary 3.2)**.** *For $M \ge n$ real polynomials of degree at most $d \ge 1$ in n real variables, the number of realisable sign assignments in $\{-1, 0, 1\}^M$ is at most*

$$(8edM/n)^n. \tag{94}$$

Zeros are counted in (94), which is why exact zero ReLU gates and exact score ties survive our enumeration. This is the multi-valued refinement of the classical bound of Warren (1968).

Theorem C.4 (van der Waerden permanent bound; Gurvits, 2005, Theorem 2.1, diagonal specialisation)**.** *If B is a $q \times q$ non-negative doubly stochastic matrix then*

$$\mathrm{per}(B) \;\ge\; q!/q^q. \tag{95}$$

The bound (95) was conjectured by van der Waerden and first proved by Falikman (1981), with the uniqueness statement due to Egorychev (1981); we quote the hyperbolic-polynomial proof of Gurvits (2005) because its diagonal specialisation yields exactly the inequality we use.

Lemma C.5 (Hilbert-space sampling; Pisier, 1981)**.** *Let $f = \sum_i v_i$ be a finite sum in a Hilbert space and $S = \sum_i \|v_i\|$. For every integer $K \ge 1$ there is an average of K rescaled atoms with squared error at most S^2/K. For $S > 0$, one may sample i with probability $\|v_i\|/S$ and use the atom $Sv_i/\|v_i\|$, omitting zero terms.*

Proof. Each sampled atom has mean f and squared norm S^2. The expected squared error of the average of K independent samples is $(S^2 - \|f\|^2)/K \le S^2/K$, so some average attains this bound. If $S = 0$, use the zero average. □

Theorem C.6 (Bounded sampling; Hoeffding, 1963)**.** *If $Y_1, \dots, Y_L$ are independent variables in $[a, b]$ with $a < b$, then for $t > 0$,*

$$\mathbb{P}\left[\frac{1}{L}\sum_{i=1}^{L}(Y_i - \mathbb{E}\,Y_i) \ge t\right] \le \exp\left(-\frac{2Lt^2}{(b-a)^2}\right). \tag{96}$$

The same bound, including the lower-tail version, holds for a uniform sample without replacement from a finite population in $[a, b]$, centred at its population mean. Applied to cell indicators, it yields the hypergeometric estimate (70).

Proposition C.7 (Finite-group character identities; Etingof et al., 2011, Sections 3.1–3.5)**.** *For a finite abelian group G, its unit-modulus characters form a group $\widehat{G}$ of order $|G|$, satisfy $\chi(u+v) = \chi(u)\chi(v)$, and obey $\sum_{\chi\in\widehat{G}} \chi(g) = |G|\mathbf{1}\{g = 0\}$. For an arbitrary finite group, with irreducible unitary representations π of dimensions d_π and characters χ_π, the regular representation identity is*

$$\sum_{\pi\in\widehat{G}} d_\pi^2 = |G|, \qquad \sum_{\pi\in\widehat{G}} d_\pi \chi_\pi(g) = |G|\mathbf{1}\{g = e\}. \tag{97}$$

D Experimental Protocols and Supplementary Results

This appendix specifies the finite-training protocol, gives the complete run-level results behind Figures 2 and 3, and records the supplementary analyses and controls; the group-family runs of Table 1 are part of the order-64 structure sweep audited in Appendix D.7. The experiments train the model class of the theory, the pair-embedding one-hidden-layer ReLU network of Section 3.2 with the L_2-penalised biases of Appendix B.2, for a fixed number of updates and report one predictor per run; the theory concerns the complete minimum-norm endpoint set with the dimensions and mask law specified in Section 3.

D.1 Protocol

Model and optimisation. The network ("PairMLP", named for its pair embedding) has separate trainable row and column embedding tables E_U, E_V in $\mathbb{R}^{128}$, one width-512 ReLU hidden layer acting on the embedded pair, and a linear q-class head, with biases on the hidden and output layers; this is the architecture of (2) with the biases of (21). Its dimensions are fixed across orders, whereas the theory uses $d_q = 2q + 1$ and $h_q = q^2$; the width-independent notch bound still requires a feasible carrier to bound endpoint costs (Appendix E.1). The common design that shares one embedding table and concatenates the two embedded inputs is the special case $E_U = W_1E$, $E_V = W_2E$ of the pair embedding. Cross-entropy is minimised by full-batch momentum SGD for 80,000 updates (learning rate 0.03, momentum 0.9, weight decay 0.001 on embeddings, weights, and biases) in float64.

Masks. The $q = 97$ orbit comparison uses one uniform random fixed-cardinality split with 2823 observed cells out of 9409; the implementation rounds $0.30q^2$, while (1) uses a floor. The $q = 64$ experiments use label-stratified masks, with the same number of observed cells from each colour. The realised sizes are

$$\begin{aligned} 832/4096 &= 20.3125\% \quad \text{(nominal 20\%)},\\ 1216/4096 &= 29.6875\% \quad \text{(nominal 30\%)},\\ 1664/4096 &= 40.625\% \quad \text{(nominal 40\%)}. \end{aligned} \tag{98}$$

For the coverage experiment, each colour has one random ordering and the masks are prefixes of that ordering, so the three levels are nested.

Seeds and reporting units. A task seed selects a table, product residual, or coordinate relabelling; a training seed selects the initialisation and optimisation repeat. Every point in the main figures is one training run; saturated markers and lines report medians.

Relative transition. The transition statistic locates sustained training memorisation, estimates a pre-generalisation test plateau and a stable terminal plateau, and counts a run when its test accuracy crosses the midpoint of that run's plateau-relative gain. It permits terminal accuracy below one; absolute endpoint accuracy is reported separately, and 0.99 is a reporting cutoff unrelated to the threshold τ.

D.2 Addition coverage across orders

The archived addition experiments vary the order q and the observed fraction α. At each order we sweep α on a grid of resolution 0.01 and record $\widehat{\alpha}_q$, the smallest tested fraction at which all five optimisation seeds reach exact held-out accuracy 1.000 under the protocol of Appendix D.1. Each grid point costs one training run per seed, so the sweep covers four orders, $q \in \{47, 79, 113, 211\}$; Table 3 is the complete record.

Table 3: Empirical observation threshold for modular addition. $\widehat{\alpha}_q$ is the smallest fraction on a grid of resolution 0.01 at which all five optimisation seeds reach exact held-out accuracy 1.000. Cell counts are the nominal $\lfloor \widehat{\alpha}_q q^2 \rfloor$ of (1); the implementation rounds, so realised sizes can differ by one cell.

order q	47	79	113	211
$\widehat{\alpha}_q$	0.33	0.25	0.23	0.20
grid cells q^2	2209	6241	12769	44521
observed cells $\lfloor \widehat{\alpha}_q q^2 \rfloor$	728	1560	2936	8904

Table 4: Reduced-polynomial reconstruction at $q = 97$. Support counts are out of $97^2 = 9409$ coefficients.

representation	tasks	nonzero coefficients	$(\deg_X, \deg_Y)$	total degree
visible addition	1	2	$(1, 1)$	1
hidden addition	10	9108–9134	$(95, 95)$	190
sequential Latin	8	9107–9127	$(95, 95)$	190

The measured fraction falls from 0.33 at $q = 47$ to 0.20 at $q = 211$ while the grid grows by a factor of about twenty. Four points at resolution 0.01, with successive decrements of 0.08, 0.02, and 0.03, fix a decline over the tested window and nothing about its shape or its monotonicity; the all-seeds criterion is one-sided, so each entry is an upper bound on the fraction at which this protocol typically succeeds. The quantity measured is a finite-training threshold under one optimiser, horizon, and architecture; the sequence α_q of Remark 4.7 is an existence statement about the full minimum-norm endpoint set with no rate. The table is not used in any proof.

D.3 The order-97 orbit comparison

The hidden-addition experiment samples ten independent triples of row, column, and symbol permutations of one cyclic-addition table. Each representation is trained under three optimisation seeds, giving 30 runs. The Latin control uses eight task seeds from the sequential random-perfect-matching generator and three optimisation seeds, giving 24 runs. Two further planned control seeds were never completed and are excluded. A one-way random-effects decomposition of the transition midpoint over the ten task seeds with three optimisation seeds each attributes about 7% of the variance to the task seed and about 93% to the optimisation seed (variance-components estimate; the fixed-effects η^2 is 0.36, and the task-seed effect is not significant, $F(9, 20) = 1.22$, $p \approx 0.33$; between-task medians span $15{,}000$–$20{,}000$ updates, mean within-task range $3{,}500$). Midpoints are read on the $1{,}000$-update checkpoint grid.

For the polynomial statistics illustrated in Figure 1, we reconstruct the unique reduced polynomial over $\mathbb{F}_{97}$ by inverting the two Vandermonde evaluation maps. Table 4 reports the result. The hidden representatives occupy the same colouring orbit as $X + Y$, while their observed-coordinate support is as dense as that of the sequential Latin controls.

D.3.1 Machine-readable coefficient archive

For every unique task instance in the order-97 comparison, we archive the unique reduced coefficient matrix over $\mathbb{F}_{97}$ once per task seed. Optimisation repeats share the same target table and therefore do not duplicate the coefficient file. Each sparse CSV contains the triples (i, j, a_{ij}) for all nonzero terms $a_{ij} X^i Y^j$, sorted first by i and then by j. The archive is `materials/polynomial-coefficients/` in the submission bundle. It also contains `manifest.csv`, which records the source experiment, support statistics, degree statistics, relative file path, and full SHA-256 digest. The 19 archived task instances comprise visible addition, ten hidden-addition relabellings, and eight sequential-Latin controls.

The median hidden-addition support is $9124/9409 = 96.971\%$. Its 30 endpoint accuracies have median 1.000 and range 0.969025–1.000; all 30 runs meet the relative-transition criterion. The 24 sequential-Latin endpoints have median 4.56×10^{-4} and range 0–1.37×10^{-3}, with 0/24 relative transitions. Both constructions are exactly Latin; the sequential generator defines a reproducible ensemble distinct from $\mathrm{Unif}(\mathcal{L}_q)$.

Table 5: Index of the machine-readable sparse coefficient archives. Each CSV stores the nonzero triples (i, j, a_{ij}) over $\mathbb{F}_{97}$.

family	seed	nonzero	support (%)	degrees	total	file
visible addition	0	2	0.021	$(1, 1)$	1	`visible_addition_task_000.csv`
hidden addition	0	9118	96.907	$(95, 95)$	190	`hidden_addition_task_000.csv`
hidden addition	1	9133	97.067	$(95, 95)$	190	`hidden_addition_task_001.csv`
hidden addition	2	9121	96.939	$(95, 95)$	190	`hidden_addition_task_002.csv`
hidden addition	3	9109	96.812	$(95, 95)$	190	`hidden_addition_task_003.csv`
hidden addition	4	9134	97.077	$(95, 95)$	190	`hidden_addition_task_004.csv`
hidden addition	5	9126	96.992	$(95, 95)$	190	`hidden_addition_task_005.csv`
hidden addition	6	9131	97.045	$(95, 95)$	190	`hidden_addition_task_006.csv`
hidden addition	7	9127	97.003	$(95, 95)$	190	`hidden_addition_task_007.csv`
hidden addition	8	9122	96.950	$(95, 95)$	190	`hidden_addition_task_008.csv`
hidden addition	9	9108	96.801	$(95, 95)$	190	`hidden_addition_task_009.csv`
sequential Latin	0	9121	96.939	$(95, 95)$	190	`sequential_latin_task_000.csv`
sequential Latin	1	9115	96.875	$(95, 95)$	190	`sequential_latin_task_001.csv`
sequential Latin	2	9107	96.790	$(95, 95)$	190	`sequential_latin_task_002.csv`
sequential Latin	3	9118	96.907	$(95, 95)$	190	`sequential_latin_task_003.csv`
sequential Latin	4	9119	96.918	$(95, 95)$	190	`sequential_latin_task_004.csv`
sequential Latin	5	9125	96.982	$(95, 95)$	190	`sequential_latin_task_005.csv`
sequential Latin	6	9127	97.003	$(95, 95)$	190	`sequential_latin_task_006.csv`
sequential Latin	7	9127	97.003	$(95, 95)$	190	`sequential_latin_task_007.csv`

Table 6: Final medians at nominal 30% label-stratified coverage; six runs per family.

family	exact accuracy	prespecified binary factor quotient
$C_2 \times C_{32}$	0.9094	0.9977
$C_2 \times C_4 \times R_8$	0.5938	0.9922
$C_2^2 \times R_{16}$	0.0165	0.9823
$C_2 \times R_{32}$	0.00330	0.9816
R_{64}	0.00069	—

D.4 Complete endpoint summaries for the open-region experiments

The designated-factor experiment (left panel of Figure 3) uses five $q = 64$ families, three task seeds per family, and two training seeds per task. Table 6 lists the final exact accuracy and the prespecified binary factor quotient. The R_{64} family has no designated binary component.

The run-level binary factor-quotient ranges for the two hardest structured families are 0.456–0.989 for $C_2^2 \times R_{16}$ and 0.478–0.984 for $C_2 \times R_{32}$. This variation motivates the run-level markers in the left panel of Figure 3.

Table 7 gives all median resolutions used in the matched nested-mask comparison. Every entry aggregates the same two task seeds and two training seeds, giving four matched triplets per family.

D.5 Factor-quotient lattice and intervention analyses

Section 5 evaluates prespecified product-factor quotients at the final checkpoint. The supplementary mechanism audit expands the evaluation to every checkpoint and, for elementary 2-group factors, to all factor-quotient directions of a fixed order. Eighteen runs contribute nine saved checkpoints each. The audit contains 550 factor-quotient-lattice records, 148 probe/intervention records, and 32 matched random-partition controls.

Two observations refine the endpoint picture. First, different optimisation runs can select different isomorphic binary factor-quotient directions. In one $C_2^2 \times R_{16}$ run, for example, prespecified binary accuracy is 0.4559 while an alternate binary factor quotient reaches 0.9858. Factor-quotient recovery is therefore naturally evaluated over the available projection lattice. Second, factor-quotient labels are linearly readable at every one of the 17 checkpoints selected for intervention, and removing the classifier component associated with the recovered factor quotient produces a substantially larger behavioural change than matched random subspaces. Removing one ridge-probe direction often has little effect; the distributed classifier component is the more faithful intervention target.

These analyses concern representation dynamics along the trajectory; their role here is to confirm that the factor-quotient accuracies in Figure 3 reflect structured label resolution rather than the contribution of exact hits to a coarse score.

Table 7: Final held-out accuracy under nested masks. A dash denotes a factor quotient absent from the factor chain.

family	resolution	20%	30%	40%
$C_2^3 \times R_8$	binary factor quotient	0.948	0.999	1.000
	4-way factor quotient	0.846	0.998	1.000
	8-way factor quotient	0.398	0.998	1.000
	exact	0.0169	0.882	0.985
$C_2^2 \times R_{16}$	binary factor quotient	0.495	0.982	0.992
	4-way factor quotient	0.344	0.960	0.984
	exact	0.00184	0.0189	0.376

D.6 Supplementary families and negative controls

Several additional sweeps delimit the empirical interpretation.

- A product ladder from C_{64} through $C_8 \times R_8$, $C_4 \times R_{16}$, $C_2 \times R_{32}$, and R_{64} shows a gradual loss of exact completion as the sequential residual factor grows. The $C_8 \times R_8$ endpoints range from 0.0427 to 0.9924, providing a finite-order non-group high-generalisation example.
- Sparse perturbations of the cyclic table yield a finite crossover at corruption levels around 2–5%. Forty valid runs remain after removing task-seed collisions. These endpoints quantify robustness to leaving the group orbit and carry no asymptotic phase label.
- A bank of nine named polynomial rules contains 81 runs. Six families show relative transitions in all nine repeats; three cross-term variants show none. The variant names used in the experiment archive are kept separate from the theoretical term mixed.
- A second stratified split reproduces the endpoint contrast between C_{64} and R_{64} in eight additional runs.
- Under the same SGD/weight-decay protocol, sixteen Transformer runs do not reach sustained training memorisation within the horizon and are therefore not comparable endpoints.

Timing-only weight-decay and angular-clock analyses remain in the experiment archive as analyses of trajectory delay.

D.7 Completeness and reproducibility audit

The order-64 structure sweep contains 72/72 completed runs and the factor-quotient sweep contains 64/64 completed runs, all reaching $80{,}000$ updates. Stored manifests and checkpoint hashes pass the archive checks. The order-97 Latin control contributes the 24 completed runs specified in Appendix D.3; denominators use completed runs only.

Archived run identities record a base commit and a dirty-worktree flag, so exact reproduction requires the frozen source snapshot, which the submission archive preserves together with the run manifests, analysis script, endpoint CSVs, figure inputs, and the finite-field polynomial reconstruction table.

E Discussion

The first two subsections discuss the stability of the all-good proof and the relation between endpoints and training timescales. The third turns to task structure: it extracts a colouring-independent carrier criterion, proves the all-good conclusion for every finite abelian group family, and examines the necessity conjecture stated in Conjecture 5.1.

E.1 What the all-good proof rests on

The competition. Stripped of its construction, the proof of Theorem 4.1(i) is the competition (13) between a counting term and a deviation term. $\mathcal{H}_q$ is the deterministic atlas of (67): the enumeration, fixed before the mask is drawn, of every cell set on which a low-norm score table can be confidently wrong while remaining consistent on the observed cells. The exponent cN of (70) is the large-

deviation price a positive-density set pays to meet a uniform mask of density α in too few cells; against it, $\log|\mathcal{H}_q|$ counts how many harmful patterns an adversary who sees the mask may choose between. The proof succeeds because (67) gives $\log|\mathcal{H}_q| = o(N)$, so the deviation term dominates and every minimum-norm response is forced to generalise. The carrier of Lemma 4.2, the SVD split, Maurey sampling, and the sign-pattern counting exist only to make $\log|\mathcal{H}_q|$ subquadratic while keeping the atlas mask-independent. That the atlas is fixed first is why the conclusion covers the whole solution set: the adversary's freedom to choose after seeing Ω has already been paid for.

Stability in width. Lemma 4.3 quantifies over tables of *any* finite width, and the atlas dimension of (64), $D_q^{\mathrm{par}} \le 2qm + mK_0 + 5qK_1 = O(q^{3/2} + r_q B_q^2/\sqrt{q} + q)$, contains no width variable: $m = \lceil\sqrt{q}\rceil$ and the atom budgets K_0, K_1 of (54) depend only on q, the norm budget B_q, and the schedule r_q. Maurey sampling replaces a sum over arbitrarily many hidden units by K_0 sampled atoms at a controlled error, so width enters the argument only through the product $\|A\|_F\|Z\|_*$. Once the width suffices to realise the carrier ($O(\log q)$ units, padded to N in the proof), further width changes the reachable norm but not the counting.

Fragility in depth. There is no depth-L analogue, and the obstruction is structural: all three ingredients of (67) are tied to a single hidden layer. The variational identity behind J reduces a two-factor L_2 objective to $\|A\|_F\|Z\|_*$, with no comparable surrogate for a composition of L factors; Maurey sampling compresses a sum of atoms, whereas depth composes them, so the Gram bound (49) that caps each normalised gate at $2q$ does not apply to inner layers; and the sign-pattern count treats preactivations of degree 2 in the core parameters, whereas at depth L they become piecewise polynomial of degree growing with L. Each of these degrades the entropy estimate in the same direction. This says the present proof does not extend, not that deep networks fail; a depth-robust form of (13) is the natural next target.

The empirical record on depth. Fan et al. (2024) train 12-layer MLPs, find deep networks at least as susceptible to delayed generalisation as shallow ones, and find that internal feature rank, not weight norm, tracks the transition; Manir & Rupa (2026) report a non-monotone depth effect on modular addition, with depth-4 MLPs failing to grok and depth-8 residual networks recovering generalisation. Both point to architectures in which upper layers are forced to track a representation built below, which is also the situation the atlas argument can cover: a single controlled map over a given feature table. A norm bound over a solution set and a rank statistic along a trajectory answer different questions, and a trajectory-level account of depth would likely be organised around a different quantity than the one controlled here.

E.2 Timescales, and the extent to which they decouple

Scope. The phase of (6) is a property of the solution set $\mathcal{S}_q(\mathcal{D})$ and carries no time index; every statement in Section 4.1 would read identically if the optimiser reached its endpoint immediately. The transition statistic of Appendix D.1 is a reporting device, not an object the theory predicts.

What our measurements show. The order-97 orbit comparison is a clean instrument for the separation: all ten hidden-addition tasks are relabellings of one addition table, so Theorem 3.5 makes the phase constant along the orbit, and the trained predictors have concentrated final accuracies ($30/30$ transitions, median 1.000, range 0.969–1.000). The delay is not constant: midpoints span 13,000–21,000 updates, with about 93% of their variance attributable to the optimisation seed (Appendix D.3). Within one orbit the phase is invariant and the delay is dominated by optimisation randomness. This does not establish that endpoint and delay are independent: every family we sweep either fixes the task and varies the optimiser, or varies the task and reports endpoints; none varies task structure while measuring delay under a comparable phase.

Where the coupling is visible. Gu et al. (2025) distinguish the accuracy ceiling from the training process leading to it, and their experiments measure both the ceiling and transition time. At $p = 97$ with a 30% training fraction, they replace the random split by excised training distributions—a square block $i, j \le k$ and a strip $i \le t$—and record both the accuracy ceiling after the transition and the epoch at which it occurs (Table 8). Enlarging the excision lowers the ceiling monotonically and shifts the transition epoch non-monotonically; the widest strip removes the transition. Under a random split

the same configuration transitions near epoch 5000 and reaches accuracy 1. One manipulation thus moves ceiling and delay together.

Table 8: Excised training distributions at $p = 97$, reproduced from Gu et al. (2025). U-Acc is the accuracy ceiling after the transition, GO-Epoch the epoch at which it occurs.

excision	parameter	U-Acc (%)	GO-Epoch
square $i, j \le k$	$k = 30$	99.04	1500
	$k = 40$	84.50	1700
	$k = 50$	77.39	3400
	$k = 60$	57.31	2200
strip $i \le t$	$t = 40$	89.89	1300
	$t = 50$	80.10	1600
	$t = 67$	7.56	none observed

A ceiling below one, with runs halting at different accuracies, is not evidence of the mixed phase: that phase concerns the exact minimum-norm set (Proposition B.2), and no training curve reports it.

How this sits against our framework. Both of their manipulations fall outside our hypotheses in instructive ways. The excisions keep the colouring at C_q^{add} and replace the mask law: our ν_q is the uniform fixed-cardinality law of (1), and uniformity is what makes $|H \cap \Omega|$ hypergeometric in (70) and so supplies the deviation term of (13). A square or strip excision is a mask that a positive-density harmful set can evade at no exponential cost, so the degradation of the ceiling is consistent with, and delimits, what the uniform-mask assumption does. Their second manipulation changes the rule to $f(i, j) = i^2 + ij + j^2$, whose ceiling stays below 1 under every split and on which a two-layer network stalls near 0.01 while a transformer retains partial generalisation; for fixed i that map is quadratic in j, so the rule is not Latin and lies outside $\mathcal{L}_q$. What survives the translation is the separation of the accuracy ceiling from the process leading to it. Our phase formalises an endpoint question by quantifying over the whole minimum-norm set under uniform masks; their experiments measure trained transformers under designed masks and expose the role of the training distribution. This comparison does not establish that the attainable accuracy is independent of the optimiser. The coupling between that ceiling and the delay is visible in their data and absent from ours; it is an open direction.

E.3 What the all-good proof needs from the colouring

Appendix E.1 asked what the proof is stable to as the model varies. Here we vary the colouring while keeping the bias-free model, the product norm (4), and the mask law (1) fixed. Only the carrier construction uses addition; the counting argument that turns a carrier into an all-good statement does not.

E.3.1 A colouring-independent notch and a carrier criterion

Lemma E.1 (One-sided notch for any colouring)**.** *Fix $\alpha, \beta \in (0, 1)$ and a deterministic sequence $C_q \in \mathcal{C}_q$. Let B_q satisfy $B_q^2 \log q / q^{5/2} \to 0$. With probability $1 - o(1)$ over $\Omega \sim \nu_q$, there is no finite-width table $D(x) = A\,\mathrm{ReLU}(Z\xi_x)$ with $\|A\|_F \|Z\|_* \le B_q$ such that*

(1) $D_{C_q(x)}(x) - D_a(x) \ge 0$ for every $x \in \Omega$ and $a \ne C_q(x)$; and

(2) at least βN cells admit a rival $a \ne C_q(x)$ with $D_a(x) - D_{C_q(x)}(x) \ge 1$.

The failure bound is uniform in the choice of the deterministic colouring.

Proof. In Appendix C.3.2, replace $y(x) = C_q^{\mathrm{add}}(x)$ by $y(x) = C_q(x)$. The Gram matrix in Step 0 depends only on the grid and on ξ. The low-rank split, sampling, and cover codes of Steps 1–3 use only the score table and its norm budget. In Step 4, the labels select the truth coordinate in $N(q - 1)$ polynomial comparisons, without changing their degrees or parameter counts. The resulting atlas may depend on C_q, but its cardinality bound does not, and it is fixed before the mask is drawn. Step 5 uses only the size of a set and the uniform mask law. Thus the same bounds prove the claim for each fixed colouring, uniformly in that colouring. This is not a simultaneous event over all colourings. □

Proposition E.2 (Carrier criterion). *Let $C_q \in \mathcal{C}_q$ be deterministic. Suppose that, for every sufficiently large q, it admits a full-grid carrier $F_q^\star = A_{\mathrm{car}}\,\mathrm{ReLU}(Z_{\mathrm{car}}\xi.)$ of width at most $h_q = N$, with $J_{\mathrm{car},q} := \|A_{\mathrm{car}}\|_F \|Z_{\mathrm{car}}\|_*$ and $\Gamma_q \geq 1$ satisfying*

$$\begin{aligned} 1 \leq F^\star_{q,C_q(x)}(x) - F^\star_{q,a}(x) \leq \Gamma_q \qquad (x \in X_q,\ a \neq C_q(x)), \\ \Gamma_q J_{\mathrm{car},q} = o\Big(\frac{q^{5/4}}{\sqrt{\log q}}\Big). \end{aligned} \tag{99}$$

Then for every fixed $\alpha, \tau \in (0,1)$,

$$A_q^-(C_q,\Omega) \xrightarrow{\mathrm{P}} 1, \qquad p^{\forall}_{q,\tau}(C_q) \to 1, \qquad \Pi_{q,\tau}([C_q]) \to \mathsf{v} = (1,0,0). \tag{100}$$

Proof. Zero padding makes the carrier feasible in (4) on every mask, so $J_* \leq J_{\mathrm{car},q}$. For any endpoint response F, take its width-N reduced representation from Lemma 3.2 and put $D = \Gamma_q F - F_q^\star$. On Ω, the observed unit margin and the upper carrier gap give $D_y - D_a \geq 0$, where $y = C_q(x)$. At every held-out non-strict error, some $a \neq y$ has $F_a \geq F_y$, so the lower carrier gap gives $D_a - D_y \geq 1$.

Balance the two representations by positive homogeneity so that $\|A_F\|_F = \|Z_F\|_* = \sqrt{J_*}$ and $\|A_{\mathrm{car}}\|_F = \|Z_{\mathrm{car}}\|_* = \sqrt{J_{\mathrm{car},q}}$. The stacked factors

$$\widetilde{A} = [\,\sqrt{\Gamma_q}A_F \ \ -A_{\mathrm{car}}\,], \qquad \widetilde{Z} = [\,\sqrt{\Gamma_q}Z_F;\ Z_{\mathrm{car}}\,] \tag{101}$$

represent D, and the same estimate as (14) gives

$$\begin{aligned} \|\widetilde{A}\|_F \|\widetilde{Z}\|_* &\leq \sqrt{2}\,(\Gamma_q J_* + J_{\mathrm{car},q}) \\ &\leq \sqrt{2}\,(\Gamma_q + 1) J_{\mathrm{car},q} =: B_q. \end{aligned} \tag{102}$$

Since $\Gamma_q \geq 1$, condition (99) is equivalent to $B_q^2 \log q / q^{5/2} \to 0$. The stack may exceed width N, which Lemma E.1 permits; the carrier itself must fit width N to justify the bound on J_*. An endpoint with $\mathrm{Acc}_T < \tau$ supplies more than $(1-\tau)(1-\alpha)N$ reverse-unit-gap cells. Apply Lemma E.1 with $\beta = \frac{1}{2}(1-\tau)(1-\alpha)$. Its exclusion event covers every endpoint simultaneously, proving (100). □

The proposition is a sufficient condition. Its contrapositive constrains families that fail to be all-good; it does not say that a family without such a carrier fails to generalise. In particular, the observed optimum J_* can be cheaper than a chosen full-grid construction, and a colouring-specific atlas can be smaller than our bound for the whole norm ball. The necessity direction in Conjecture 5.1 is an additional hypothesis.

E.3.2 Finite abelian group tables

Let G be a finite abelian group of order q, identify its elements with $[q]$, and write its operation additively. Its table $C_G(u,v) = u + v$ is Latin. Let $\widehat{G}$ be its character group, with q elements, and write $\chi(t) = c_\chi(t) + i s_\chi(t)$, where $c_\chi(t)^2 + s_\chi(t)^2 = 1$. We use the standard finite-group character identities of Etingof et al. (2011, Sections 3.1–3.5).

Lemma E.3 (Abelian carrier). *There is an absolute constant C_{ab} such that, for every finite abelian group G of sufficiently large order q, the table C_G has a full-grid carrier of width $O(\log q)$, truth–rival gaps in $[1,9]$, and $J_{\mathrm{car},q} \leq C_{\mathrm{ab}} q\sqrt{\log q}$.*

Proof. Keep h, ϵ, ρ, c, L and the ReLU product surrogate P from Appendix C.3.1. We give each change to that construction.

Character selection. For $e \neq 0$, character orthogonality gives $\sum_{\chi \in \widehat{G}} \chi(e) = 0$. Indeed, a character ψ with $\psi(e) \neq 1$ exists, and multiplying the sum by $\psi(e)$ merely permutes its terms. Thus, if $\mathbf{1}$ denotes the trivial character,

$$\mathbb{E}_{\chi \sim \mathrm{Unif}(\widehat{G}\setminus\{\mathbf{1}\})}[c_\chi(e)] = -\frac{1}{q-1}. \tag{103}$$

The real parts lie in $[-1, 1]$. Hoeffding's inequality and a union bound over the $q-1$ nonzero elements therefore give, with the same choice of $L = O(\log q)$, characters $\chi_1, \dots, \chi_L$ such that

$$\frac{1}{L}\sum_{\ell=1}^{L} c_{\chi_\ell}(e) \le \rho \qquad (e \neq 0). \tag{104}$$

Sampling with replacement is allowed; no distinctness is needed.

Products, scores, and gaps. Multiplicativity gives the real and imaginary addition identities. Set

$$\begin{aligned} \widetilde{c}_\ell(u,v) &= P(c_{\chi_\ell}(u), c_{\chi_\ell}(v)) - P(s_{\chi_\ell}(u), s_{\chi_\ell}(v)), \\ \widetilde{s}_\ell(u,v) &= P(s_{\chi_\ell}(u), c_{\chi_\ell}(v)) + P(c_{\chi_\ell}(u), s_{\chi_\ell}(v)). \end{aligned} \tag{105}$$

By (43), these approximate $c_{\chi_\ell}(u+v)$ and $s_{\chi_\ell}(u+v)$ to error at most $h^2/8$. Define

$$F^\star_{q,a}(u,v) = \frac{c}{L}\sum_{\ell=1}^{L}\big(c_{\chi_\ell}(a)\widetilde{c}_\ell(u,v) + s_{\chi_\ell}(a)\widetilde{s}_\ell(u,v)\big). \tag{106}$$

Its ideal version is $\frac{c}{L}\sum_\ell \operatorname{Re}\chi_\ell(u+v-a)$, which is c at the true label and at most $c\rho$ at a rival. Since $|c_\chi(a)| + |s_\chi(a)| \le \sqrt{2}$, the score error is at most $c\sqrt{2}h^2/8$. The margin calculation in Appendix C.3.1 therefore gives exactly the same interval $[1, 9]$.

Architecture and cost. Expanding P using (41) produces a constant number of hinges per character. Each hinge is $(r_u + s_v - t)_+$, with bounded row and column potentials; splitting t equally between them makes it a single reduced ReLU unit. The preactivation row has norm $O(\sqrt{q})$, and its output column has norm $O(\sqrt{q}/L)$, with constants independent of G. Consequently

$$\|A_{\mathrm{car}}\|_F = O(\sqrt{q/L}), \qquad \|Z_{\mathrm{car}}\|_* \le \sqrt{O(L)}\|Z_{\mathrm{car}}\|_F = O(L\sqrt{q}), \tag{107}$$

giving $J_{\mathrm{car},q} = O(q\sqrt{\log q})$ and width $O(\log q) \le N$. □

Proof of Corollary 4.4. Lemma E.3 gives $\Gamma_q J_{\mathrm{car},q} = O(q\sqrt{\log q}) = o(q^{5/4}/\sqrt{\log q})$. Apply Proposition E.2 and Theorem 3.5. □

Thus the all-good pole contains every abelian group family, beyond the cyclic case of Theorem 4.1(i). Non-isomorphic abelian groups give distinct colouring orbits: an isotopy between two such group tables can be written $h(u+v) = f(u) + g(v)$; substituting the identity in each argument shows that $u \mapsto h(u) - h(0)$ is a group isomorphism. Transposition changes no abelian table. At orders $q = 2^m$, partitions of m therefore supply an unbounded number of distinct abelian orbits (11 at $m = 6$), all covered by the same asymptotic bound. This is a limit statement along families, not a claim that a finite-order profile equals ∨ exactly.

E.3.3 What the finite-order experiments can suggest

The abelian rows C_2^6, C_{64}, and $C_8 \times C_8$ in Table 1 are finite members of families covered by Corollary 4.4. The corollary concerns the whole minimum-norm endpoint set under ν_q as $q \to \infty$; the table reports one trained predictor per run at $q = 64$, under label-stratified masks and a fixed horizon. It neither tests the endpoint quantifier nor orders the finite-training accuracies.

There is nevertheless a concrete simplification for elementary abelian 2-groups. Every character is real with values in $\{\pm 1\}$, so $s_\chi = 0$ and (105) reduces to one product $P(c_\chi(u), c_\chi(v))$ per character. Its two arguments have sum and difference in $\{-2, 0, 2\}$, interpolation nodes of Q_h, hence $P(x, y) = xy$ exactly on these inputs. The construction uses fewer hinges and has no multiplication error. These are improvements to this explicit circuit, not a comparison of the minimum achievable carrier norms across groups or a prediction of the $6/6$ exact completions for C_2^6.

E.3.4 NON-ABELIAN TABLES AND THE MISSING COST ESTIMATE

The transitions for Dih_{64} invite an extension beyond scalar characters. Every one-dimensional character of a non-abelian group factors through its abelianisation. It therefore equals 1 on every commutator, so no selection of such characters can satisfy (104) at a non-identity commutator. For the order-64 dihedral group, the abelianisation has only four elements.

Write $\widehat{G}$ now for the irreducible unitary dual, let $\pi \in \widehat{G}$ have dimension d_π, and set $\chi_\pi = \operatorname{tr} \pi$. The regular-representation identity gives

$$\sum_{\pi \in \widehat{G}} \frac{d_\pi^2}{q} \frac{\chi_\pi(g)}{d_\pi} = \mathbf{1}\{g = e\}. \tag{108}$$

The real part of the normalised character lies in $[-1, 1]$, so sampling with weights d_π^2/q gives a separation argument analogous to the abelian one. The candidate ideal score for label a uses $\operatorname{Re} \chi_\pi(uva^{-1})/d_\pi$, rather than $\chi_\pi(uv)$ alone. In particular,

$$\frac{\chi_\pi(uva^{-1})}{d_\pi} = \frac{1}{d_\pi} \sum_{i,j,k=1}^{d_\pi} \pi(u)_{ik} \pi(v)_{kj} \overline{\pi(a)_{ij}}. \tag{109}$$

Each complex product of a row quantity and a column quantity can again be expanded into real products and approximated using P. But this direct construction contains d_π^3 scalar complex products shared across the label scores; counting only the d_π^2 terms in $\operatorname{tr}(\pi(u)\pi(v))$ does not cost all those scores. Moreover, multiplication errors accumulate with dimension, so the mesh may need refinement, and the nuclear and output norms must be recomputed. We leave these costs open. Families with uniformly bounded irreducible dimensions, including dihedral groups, are a concrete first target for Proposition E.2; growing dimensions require quantitative control of both approximation and representation size.

E.3.5 RARITY, AND WHAT A NECESSARY CARRIER WOULD MEAN

Every colouring has a full-grid carrier of finite, indeed polynomial, cost: the cell-indicator construction in Appendix C.1 uses N units and bounded parameter entries. This includes non-group Latin tables and uniformly random Latin squares. What is unknown for the non-group products of Section 5 is a construction cheap enough to satisfy (99) along an infinite family.

The uniform obstruction does give a precise rarity statement for cheap carriers. For a deterministic $b_q = o(q^{5/4}/\sqrt{\log q})$, let $\mathcal{K}_q(b_q) \subseteq \mathcal{L}_q$ contain the colourings with a width-at-most-N carrier whose gaps lie in $[1, \Gamma]$ and whose cost satisfies $\Gamma J_{\mathrm{car}} \le b_q$ for some $\Gamma \ge 1$. The proof of Proposition E.2, using $B_q = 2\sqrt{2} b_q$, gives $p^{\forall}_{q,\tau}(C) \ge 1 - o(1)$ uniformly on this set. Thus Theorem 4.1(ii) implies

$$\frac{|\mathcal{K}_q(b_q)|}{|\mathcal{L}_q|} (1 - o(1)) \le P^{\forall}_{q,\tau}(\mathsf{M}_{\mathrm{unif}}) \longrightarrow 0. \tag{110}$$

This rules out this cheap-carrier regime for almost all Latin colourings; it rules out neither arbitrary carriers nor the existence of a good endpoint in a mixed instance.

Why the earlier quantitative converse is too strong. The sufficient budget in (99) is not necessary. The accompanying *Carrier necessity counterexample* note constructs a deterministic Latin family along even orders for which $A_q^- \xrightarrow{\mathrm{P}} 1$ at every fixed positive observation fraction, yet the minimum of ΓJ_{car} over all width-at-most-N carriers with gaps in $[1, \Gamma]$ is $(1 + o(1))q^{13/10}$. Its ratio to $q^{5/4}/\sqrt{\log q}$ tends to infinity. The construction uses independent four-cell Latin trades and the first crossing of a carrier-cost threshold; a multiscale margin count still establishes all-goodness. Thus the counterexample separates the old numerical budget from the broader carrier mechanism. Its lower bound concerns the joint cost ΓJ, not the unit-margin cost without Γ.

A mask-independent robust certificate. For a colouring C and $0 \le \varepsilon < 1$, define

$$\mathcal{R}_q(C, \varepsilon) = \inf \left\{ \|A\|_F \|Z\|_* : \begin{array}{l} w \le N, \quad E \subseteq X_q, \quad |E| \le \varepsilon N, \\ F_{C(x)}(x) - F_a(x) \ge 1 \quad (x \notin E,\ a \ne C(x)) \end{array} \right\}, \tag{111}$$

where $F = A\,\mathrm{ReLU}(Z\xi.)$. The infimum ranges jointly over the representation and the exceptional set E. Both are chosen from the deterministic task, independently of the mask. No upper gap is prescribed, and the architecture and width restriction remain those of Section 3.

In this notation, Conjecture 5.1 asserts that if a deterministic Latin sequence satisfies $A_q^-(C_q, \Omega) \xrightarrow{\mathrm{P}} 1$ for every fixed $\alpha \in (0, 1)$ under its corresponding mask law, then there exists a deterministic sequence $\varepsilon_q \to 0$ such that

$$\mathcal{R}_q(C_q, \varepsilon_q) = o(q^{3/2}). \tag{112}$$

The premise includes every fixed accuracy threshold $\tau < 1$, separately at each fixed observation fraction; it does not assert a single event uniform in α. The conclusion is equivalent to the existence of the carriers in the main-text statement: near-minimisers of the infimum suffice to select their representations and exceptional sets.

What the revision changes. The upper-gap factor Γ pays for the subtraction $\Gamma F - F_q^\star$ in the original proof, whereas a direct margin argument can work with F itself. Removing it avoids treating that proof cost as an intrinsic necessary complexity. Allowing $o(N)$ exceptional cells matches an accuracy limit that can ignore a vanishing fraction of difficult cells. Requiring all-goodness at every fixed positive observation fraction avoids assuming an unproved transfer from one fraction to all others. These are independent changes to the converse, not a claim that merely replacing the exponent $5/4$ by $3/2$ gives a sharp characterisation.

Why necessity remains open. The accompanying multiscale exploration gives a sufficient condition using a full-grid unit-margin carrier of cost $o(q^{3/2}/\log q)$, without an upper-gap factor. It does not prove (112): all-goodness supplies neither an upper bound on J_* nor a quantitative lower bound on the positive held-out margins. Scaling a good endpoint to unit margin outside a small exceptional set may therefore have uncontrolled cost. Averaging endpoints over masks does not resolve this problem, since accuracy does not control the magnitude of rare negative gaps or the norm of a combined representation. An entropy upper bound also cannot be reversed to force bad global minimisers above its sufficient budget.

Conversely, a robust carrier need not satisfy the constraints at observed exceptional cells, so it does not by itself give $J_* \le \mathcal{R}_q(C_q, \varepsilon_q)$. Repairing an arbitrary $o(N)$-cell set requires additional cost and width control, beyond the four-cell trade argument. The revised conjecture is therefore only a proposed necessary condition, not an equivalence; neither necessity of the exponent $3/2$ nor a critical logarithmic factor is established.