# Compositional Generalization Certificates via the Van Kampen Theorem

**Karen Sargsyan**
Institute of Chemistry, Academia Sinica
karen.sarkisyan@gmail.com

## Abstract

A neural network trained on the multiplication table of the quaternion group $Q_8$ can achieve perfect accuracy within each of two overlapping cyclic subgroups and predict every cross-region product wrong. In the quaternion example, exact local representations that agree on the shared subgroup but differ there from a fixed faithful target have matrix loss four in complex dimension two. We align local representations on their finite overlap by a chosen unitary basis change. The van Kampen universal property then extends the aligned pair uniquely to the amalgamated free product. This gluing is described by a homotopy pullback of representation groupoids. The resulting certificate bounds cross-region matrix error using local approximation errors, agreement with the target, overlap structure, and how faithfully the model composes its operators. For central cyclic overlap, we show the lower bound $2(d - r_{\max})$ under uniform sampling on complete overlap orbits, where $r_{\max}$ counts matching overlap eigenvalues. This provides a quantitative framework for assessing generalization to unseen compositions, showing what locally consistent operators guarantee and what additional conditions are needed for accurate prediction.

## 1 Introduction

This paper concerns a failure of compositional generalization: a model trained on multiplication in the quaternion group $Q_8$ can achieve perfect within-region accuracy on $\langle i\rangle = \{1, i, -1, -i\}$ and $\langle j\rangle = \{1, j, -1, -j\}$ while predicting every cross-region product wrong. These are the cyclic subgroups generated by the quaternion elements $i$ and $j$; their overlap is $C = \{1, -1\}$.

We analyze the corresponding matrix predictions using unitary representations, which assign unitary matrices to group elements and preserve multiplication. In dimension $d = 2$, each exact, compatible local predictor that represents the shared element $-1$ by a matrix other than the negative identity has squared Frobenius loss exactly $4$ against a fixed faithful target (one that distinguishes all group elements) on four cross pairs. This number is computable before any training.

The framework that produces this number—a stratified generalization certificate built from the van Kampen theorem—does more than explain the $Q_8$ case. It constructs a consistent predictor from compatible local representations and bounds its separation from a specified target. For central cyclic overlaps, whose elements commute with both factors, character orthogonality gives a general lower bound. A character records the trace of each representation matrix. The quaternion example shows when that bound is sharp and when the fixed target forces a larger error.

**Setup.** Modular AI systems are built from components that share inputs. A vision module and a language module share an embedding space; two experts in a mixture share a routing variable; two subnetworks in a multi-task system share a latent bottleneck. The ability to combine locally correct modules into globally correct predictions—*compositional generalization* [4, 10], of which

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

group-composition learning is the algebraic incarnation [16, 12, 14]—is a central challenge for neural networks. Group multiplication tasks are the simplest mathematical incarnation: the network learns the multiplication table of a group $G$ presented as $G = \langle A, B\rangle$ for two subgroups $A, B \leq G$, trains only on within-region pairs (both factors in $A$ or both in $B$), and is tested on cross-region pairs. The overlap $C = A \cap B$ is the shared variable that both modules must handle consistently.

**Mechanism.** The set $\mathrm{Hom}(G, \mathrm{U}(d))$ of $d$-dimensional unitary representations of a group $G$ is a topological space; $\mathrm{U}(d)$ denotes the unitary group. The amalgamated free product $\widetilde{G} = A *_C B$ records the rules within $A$ and $B$, identifying their common elements. The universal property underlying the van Kampen theorem [22] extends a specified compatible pair uniquely to $\widetilde{G}$. The representation-groupoid formulation, which keeps track of representations and their unitary basis changes, expresses this gluing as a homotopy pullback (Appendix G). Concretely: the gluing data consist of a representation of $A$, a representation of $B$, and a chosen unitary basis change (an intertwiner) aligning their two restrictions on $C$.

The strict pullback would demand exact matrix equality of the two restrictions—same matrices in the same basis. The homotopy pullback allows agreement up to conjugation—same representation, possibly different basis—and records the chosen conjugating unitary. For learning, where independently trained modules may use different internal coordinates, this formulation records the alignment needed to compose their operators.

For finite $C$, connected components of $\mathrm{Hom}(C, \mathrm{U}(d))$ are unitary conjugacy classes, the *strata* or overlap types; representations in different strata cannot be aligned by any unitary conjugation. The task group $G$ may impose further relations: in the quaternion case, the local laws $i^2 = j^2 = -1$ do not impose $ji = -ij$.

Two failures are now distinguishable. The local representations may have different overlap types, preventing alignment. Or they may agree with each other but share a type different from the target's. For central cyclic overlap, a matrix-error gap appears when evaluation uniformly samples the pairs obtained by multiplying the first input of each base pair by every overlap element (complete overlap-orbit sampling). Even the correct type leaves choices within a stratum, so target accuracy requires more than type agreement.

**Connection to learning.** Group-composition tasks provide controlled settings for grokking, where generalization improves well after training accuracy becomes high [16, 12]. The certificate concerns a fixed model: what do its learned operators imply about unobserved compositions? For a classifier, this requires checking both the operator fit and its connection to the output. Section 6 gives measurable conditions and a character probe based on fitted operator traces; the frozen-arm experiment separates matrix error from label accuracy. Training trajectories and the timing of transitions remain separate questions.

**Related work.** A rich body of work analyzes group-composition learning. Mechanistic interpretability reveals what grokked networks learn: Fourier features aligned with group irreducible representations (irreps) [15, 3], with grokking linked to the lazy-to-rich transition [8] and Fourier structure [5]. In their two-layer setting, Marchetti et al. [14] prove that networks learn irreps one at a time in Fourier-statistic order, and show how deeper architectures exploit associativity.

On the compositional side, Lippl and Stachenfeld [11] also study held-out compositional splits, deriving a *conjunction-wise additivity* constraint for kernel models with fixed, compositionally structured representations. Their restriction concerns sums of contributions from component combinations seen in training; ours concerns products of local group operators and the representation carried by their shared subgroup. Ram et al. [17] study compositional complexity and its implications for expressivity and generalization; Wang and Wang [20] develop an algebraic and geometric account of how training can yield symbolic structure. Provable compositional generalization has also been obtained under assumptions on additive decoders [9] and architectures with compositional structure [23]. Abbe et al. [1] study generalization to unseen Boolean combinations through degree and the Fourier–Walsh basis. Functorial architectures built from higher inductive types [18] provide a complementary route by imposing compositional structure in the architecture; the present construction certifies operator families satisfying explicit measured hypotheses.

**Contributions.** We bring a tool from algebraic topology—the van Kampen theorem—to prove a *stratified generalization certificate*.

**1. The compositional predictor via universal property.** Given approximate local representations, we correct them, align them on the overlap, and extend the resulting pair uniquely to $A *_C B$ (Section 3). The certificate bounds the approximate product's distance from this exact extension.

**2. Topological obstruction.** The overlap stratum supplies a discrete invariant. For central cyclic overlap and complete overlap-orbit sampling, a mismatch with the target gives a positive separation. This transfers to model outputs with a controlled perturbation and output-disagreement budget.

**3. Exact quaternion separation and a general bound.** For $Q_8$ in dimension two, we compute both the fixed-target and relaxed-target gaps. For central cyclic overlaps we prove the lower bound $2(d - r_{\max})$ by character orthogonality, where $r_{\max}$ counts matching overlap eigenvalues (Sections 4–5).

**4. Conditions for using the certificate.** We give measurable operator-extraction and alignment conditions, identify limits of the character probe, and illustrate the separation with frozen representations and trained readouts (Sections 6–7).

Readers unfamiliar with finite-group representation theory may wish to consult Appendix A, which collects mathematical preliminaries on groups, representations, and $\mathrm{Hom}(C, \mathrm{U}(d))$. Appendix I gives an informal walkthrough using two clocks sharing an overlap. Detailed proofs are in the intervening appendices.

## 2 Setup and Definitions

We use finite groups $A, B \leq G$ with $G = \langle A, B \rangle$ and $C = A \cap B$. Write $\widetilde{G} = A *_C B$ and let $\pi : \widetilde{G} \twoheadrightarrow G$ be the natural quotient. The unitary group is $\mathrm{U}(d)$; $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_F$ denote operator and Frobenius norms. All matrix comparisons use one specified coordinate system and inner product. Squared Frobenius losses are not divided by the dimension.

*Remark* 1 (Universal property and alignment). The amalgam is the pushout in the category of groups. Thus

$$\mathrm{Hom}(\widetilde{G}, \mathrm{U}(d)) \cong \mathrm{Hom}(A, \mathrm{U}(d)) \times_{\mathrm{Hom}(C, \mathrm{U}(d))} \mathrm{Hom}(B, \mathrm{U}(d)).$$

This is a strict pullback of representation sets: the restrictions agree as matrices. If they initially agree only up to conjugation, one first chooses a conjugating unitary and changes one local basis. The representation-groupoid formulation records that unitary as part of the gluing data (Appendix G).

If a network has learned local multiplication tables $\rho_A$ and $\rho_B$ that agree on the shared elements $C$, what is the natural prediction for a cross-region product $a \cdot b$? Simply multiply: $\rho_A(a) \cdot \rho_B(b)$. This is the compositional predictor.

**Definition 2** (Compositional predictor)**.** For exact representations $\rho_A : A \to \mathrm{U}(d)$ and $\rho_B : B \to \mathrm{U}(d)$ agreeing on $C$, let $\widetilde{\rho}$ be their unique extension to $\widetilde{G}$. For $(a, b) \in A \times B$, define $\widehat{y}(a, b) = \rho_A(a)\rho_B(b) = \widetilde{\rho}(ab)$.

The extension descends to $G$ precisely when it is trivial on $\ker \pi$. This involves the missing task relations, not just the overlap type. Local multiplication data describe compatible extensions to the amalgam; a particular model specifies one such extension only after its local operators have been fixed.

**Definition 3** (Overlap type and strata)**.** The overlap type of a representation is the unitary conjugacy class of its restriction to $C$. For finite $C$, these classes are exactly the connected components of $\mathrm{Hom}(C, \mathrm{U}(d))$. For $C = \{e, z\} \cong \mathbb{Z}/2$, the type is the multiplicity $m \in \{0, \ldots, d\}$ of $-1$ in $\rho(z)$.

We write $\sigma(\rho)$ for the overlap type of $\rho$. Only types that extend to both factors are admissible. In the quaternion amalgam all three types $m = 0, 1, 2$ are realizable in dimension two, although actual $Q_8$ representations in that dimension have only $m = 0$ or $m = 2$.

What does it mean for a network to "approximately know" a multiplication table? Not that it matches a specific target, but that its outputs are internally consistent—nearly multiplicative.

**Definition 4** (Local quasi-representations)**.** For $H = A, B$, let $F_H : H \to \mathrm{U}(d)$ satisfy $F_H(e) = I_d$, and set

$$\varepsilon_H = \max_{g,h\in H} \|F_H(gh) - F_H(g)F_H(h)\|_{\mathrm{op}}, \qquad \zeta = \max_{c\in C} \|F_A(c) - F_B(c)\|_{\mathrm{op}}.$$

The operator prediction is $S(a, b) = F_A(a)F_B(b)$.

The two maps allow independently fitted modules. When they restrict one common $F_\theta$, $\zeta = 0$. Small $\varepsilon_H$ means local self-consistency; it does not measure agreement with a target. Nonunitary fits require correction using the unitary factor of the polar decomposition, treated in Section 6.

**Definition 5** (Target loss and separation)**.** Fix $\rho_t \in \mathrm{Hom}(G, \mathrm{U}(d))$ and a probability distribution $\mu$ on evaluated pairs in $A \times B$. For a matrix output $Z$, write

$$\mathcal{L}(Z) = \mathbb{E}_\mu \|Z(a, b) - \rho_t(ab)\|_F^2.$$

For a realizable type $\sigma$, the fixed-target separation is

$$c_{\rho_t}(\sigma; \mu) = \inf_{\substack{\rho\in\mathrm{Hom}(\widetilde{G},\mathrm{U}(d))\\ \sigma(\rho)=\sigma}} \mathbb{E}_\mu \|\rho(ab) - \rho_t(ab)\|_F^2. \tag{1}$$

The relaxed separation $c_{\mathrm{rel}}(\sigma, \sigma_t; \mu)$ also minimizes over the target representation of $\widetilde{G}$ with type $\sigma_t$. Hence $c_{\mathrm{rel}} \le c_{\rho_t}$ when $\sigma_t = \sigma(\rho_t \circ \pi)$.

These infima are attained for finite factors and a realizable type, by compactness (Appendix A). Different types need not have positive separation for an arbitrary $\mu$.

### 2.1 Underdetermination and representation obstructions

Local multiplication laws and target accuracy concern different objects. The laws within $A$ and $B$, together with their shared elements, define the amalgam $\widetilde{G}$. The task additionally specifies the quotient $G$. An exact compatible operator family respects the local laws, but respects the task's remaining relations only if it belongs to

$$\{\rho \in \mathrm{Hom}(\widetilde{G}, \mathrm{U}(d)) : \ker \pi \subseteq \ker \rho\} \cong \mathrm{Hom}(G, \mathrm{U}(d)).$$

Thus local consistency alone leaves a choice of extensions. Even among extensions that descend to $G$, comparison with a fixed matrix target requires agreement with that target in the specified coordinates.

The overlap type describes a particular operator family within this space of possibilities. It is coarser than the extension itself: many different factor representations can have the same restriction type. A positive separation $c_{\rho_t}(\sigma; \mu)$ states that every exact extension in a specified type incurs a minimum error on the evaluated products. This is a restriction on that family of predictors, beyond the observation that local data leave the target underdetermined. The perturbation bound below transfers the restriction to approximate operators. Appendix J illustrates underdetermination with a connected infinite-overlap representation space.

The quaternion task exhibits both effects within the same setting. The shared relation $x^2 = y^2$ leaves the relative position of the two generators unspecified. The target imposes $yx = x^{-1}y$ as well. Section 4 shows that the two wrong overlap types have fixed-target loss four, while the correct type contains a continuous family of losses, including zero. Consequently, identifying the type can rule out accurate matrix prediction without uniquely identifying the predictor. Its force depends on the target and evaluated pairs: Section 5 gives a wrong-type predictor that is exact on a single cross pair but fails on its complete overlap orbit.

## 3 The Stratified Generalization Certificate

### 3.1 Correction and alignment

The correction step belongs to the Hyers–Ulam stability tradition [21, 6]; Kazhdan's work on approximate unitary representations [7] supplies the stability result used here for finite groups. Suppose exact local corrections $\widehat{\rho}_H : H \to \mathrm{U}(d)$ have errors

$$\alpha_H = \max_{h\in H} \|F_H(h) - \widehat{\rho}_H(h)\|_{\mathrm{op}}. \tag{2}$$

For normalized maps on finite groups, the stability bound of Burger et al. [2, Theorem 1.1] gives $\alpha_H \le \varepsilon_H + 120\varepsilon_H^2$ when $\varepsilon_H < 1/10$. We keep the correction errors explicit, so supplied corrections can also be checked directly.

**Lemma 6** (Finite-overlap alignment)**.** *For exact $\sigma_1, \sigma_2 : C \to \mathrm{U}(d)$ with $\tau = \max_{c\in C} \|\sigma_1(c) - \sigma_2(c)\|_{\mathrm{op}} < 1$, there is a unitary $u$ such that $u\sigma_1(c)u^* = \sigma_2(c)$ for every $c \in C$ and $\|u - I_d\|_{\mathrm{op}} \le 2\tau$.*

The construction averages $\sigma_2(c)\sigma_1(c)^*$ over $C$ and takes the polar unitary. Thus alignment has a direct finite formula. The complete proof is in Appendix B.

### 3.2 Main theorem

The theorem has three components: a predictor construction, an upper bound from local target agreement, and a lower bound from stratum separation.

**Theorem 7** (Conditional stratified certificate)**.** *Suppose the corrections in* (2) *satisfy*

$$\tau = \max_{c\in C} \|\widehat{\rho}_A(c) - \widehat{\rho}_B(c)\|_{\mathrm{op}} < 1.$$

*Choose $u$ from Lemma 6, and set $\rho_A = u\widehat{\rho}_A u^*$, $\rho_B = \widehat{\rho}_B$, $\ell = \|u - I_d\|_{\mathrm{op}}$, and $\delta = \alpha_A + \alpha_B + 2\ell$. Then the following statements hold.*

***Construction.*** *The aligned pair has a unique extension $\widetilde{\rho}$ to $\widetilde{G}$, with predictor $\widehat{y}$ and overlap type $\sigma_F$. Moreover,*

$$\max_{a,b} \|S(a,b) - \widehat{y}(a,b)\|_{\mathrm{op}} \le \delta, \qquad \delta \le \alpha_A + \alpha_B + 4\tau. \tag{3}$$

***Target comparison.*** *For any matrix output $Z$, let $\kappa = (\mathbb{E}_\mu \|Z - S\|_F^2)^{1/2}$ and $\eta_H = \max_{h\in H} \|F_H(h) - \rho_t(h)\|_{\mathrm{op}}$. With $[x]_+ = \max(x, 0)$,*

$$\mathcal{L}(Z) \ge [\sqrt{c_{\rho_t}(\sigma_F; \mu)} - \sqrt{d}\,\delta - \kappa]_+^2, \tag{4}$$

$$\mathcal{L}(Z) \le [\sqrt{d}\,(\eta_A + \eta_B) + \kappa]^2. \tag{5}$$

*For the actual operator product $Z = S$, $\kappa = 0$.*

**Corollary 8** (A sufficient alignment condition)**.** *If $\zeta + \alpha_A + \alpha_B < 1$, the theorem applies and $\delta \le 5(\alpha_A + \alpha_B) + 4\zeta$. If the corrections already agree on $C$, take $u = I_d$ and $\delta = \alpha_A + \alpha_B$.*

**Proof sketch.** **Why the predictor exists.** The first step rounds each approximately consistent local multiplication table to an exactly consistent one (Ulam stability). The two corrected tables, one for $A$ and one for $B$, may use different internal coordinate systems. After aligning that basis, the two tables agree exactly. The van Kampen universal property then says: there is exactly one way to extend these aligned local tables to a representation of the amalgam. This is the step where van Kampen does essential work. Once the tables and alignment are fixed, the rest of the argument is a perturbation around that candidate.

**Why the upper bound holds.** If each local operator is close to its target, their product is close to the target product, with error bounded by the sum of the two local errors. This is independent of the stratum. The theorem also accounts for disagreement between the operator product and the model's output.

**Why the lower bound holds.** The predictor $\widehat{y}$ is a legitimate representation of $\widetilde{G}$ in stratum $\sigma_F$. If it is separated from the target on the tested products, small local and output errors preserve part of that gap. The quaternion and cyclic calculations below supply the separation; the reverse triangle inequality transfers it to the output. The full proof is in Appendix C.

## 4 The Exact Quaternion Separation

The heart of the obstruction is computing the stratum separation explicitly. For $Q_8$, this reduces to a clean sign-pattern argument.

The amalgam is $\widetilde{G} = \langle x, y \mid x^4 = y^4 = 1,\ x^2 = y^2 \rangle$; the quotient $Q_8$ also imposes $yx = x^{-1}y$. Fix a faithful target $\rho_t : Q_8 \to \mathrm{U}(2)$ and write $I_t = \rho_t(i)$, $J_t = \rho_t(j)$, $K_t = I_t J_t$. Let $\mu$ be uniform

on $(i,j),(i,-j),(-i,j),(-i,-j)$. For an amalgam representation, set $X=\rho(x)$, $Y=\rho(y)$, $M=X^2=Y^2$, and $P=XY$. Then $M$ commutes with $X,Y$ and $M^2=I$.

The sign-pattern table makes the mechanism transparent:

| Cross pair | Local predictor | Faithful target |
|---|---|---|
| $(i,j)$ | $P$ | $K_t$ |
| $(i,-j)$ | $MP$ | $-K_t$ |
| $(-i,j)$ | $MP$ | $-K_t$ |
| $(-i,-j)$ | $P$ | $K_t$ |

(6)

**Proposition 9** (Fixed and relaxed quaternion targets)**.** *In complex dimension two, for* $m=\dim\ker(M+I_2)$*,*

$$c_{\rho_t}(0;\mu)=4, \qquad c_{\rho_t}(1;\mu)=4, \qquad c_{\rho_t}(2;\mu)=0. \tag{7}$$

*Every exact predictor of type* $m=0$ *or* $m=1$ *has average loss four. In any complex dimension* $d$*, if the target instead varies over all amalgam representations with overlap* $-I_d$*, the joint infimum is*

$$c_{\mathrm{rel}}(m,d;\mu)=2(d-m). \tag{8}$$

*Thus its values in dimension two are* $4,2,0$*.*

The element $-i$ equals $(-1)\cdot i$, so the predictor outputs $\rho(-i)=\rho(-1)\rho(i)=MX$. If $M=+I_2$ ($m=0$), then $\rho(-i)=\rho(i)$—the predictor identifies $i$ and $-i$. All four cross predictions collapse to the same matrix $P$. A constant pattern cannot approximate the target's alternating one; the parallelogram identity gives the gap of four. For $m=1$, the two distinct eigenlines of $M$ are preserved by both $X$ and $Y$. Consequently $P$ and $MP$ are Hermitian, while $K_t$ is anti-Hermitian. Their real Frobenius inner products vanish, again giving four. A faithful predictor attains zero at $m=2$. Appendix D proves these statements and constructs witnesses for (8). The difference between fixed and relaxed targets is essential: their mixed-type minima are not equal.

For corrected predictors of either wrong type, Theorem 7 therefore gives

$$\mathcal{L}(Z)\geq[2-\sqrt{2}\,\delta-\kappa]_+^2. \tag{9}$$

The constant four is a matrix-space quantity in a specified dimension and norm; cross-entropy and classification error are separate quantities.

### 4.1 Local label supervision and admissible types

Local classification places an additional restriction on exact operator models. Here the decoder receives only a product matrix, with no separate access to the input labels.

**Proposition 10** (Local label separation)**.** *Let* $H=A$ *or* $B$*, let* $\rho_H: H\to \mathrm{U}(d)$ *be exact, and suppose a decoder* $D$ *satisfies* $D(\rho_H(a)\rho_H(b))=ab$ *for every* $(a,b)\in H\times H$*. Then* $\rho_H$ *is injective. In particular, if* $C=\{e,c\}$ *with* $c\neq e$*, its overlap matrix has at least one negative eigenvalue:* $m\geq 1$*.*

Indeed, the inputs $(e,h)$ require the matrices $\rho_H(h)$ to receive distinct labels. The involution $\rho_H(c)$ must therefore differ from $I_d$ and have a $-1$ eigenvalue. A full proof and a quaternion witness are given in Appendix D.

For $Q_8$ in dimension two, $m=0$ identifies $i$ with $-i$ and is excluded by perfect local classification. The mixed type is still possible: $X=\mathrm{diag}(1,\mathrm{i})$ and $Y=\mathrm{diag}(-1,\mathrm{i})$ give six distinct local matrices. A shared decoder can assign all six labels and classify all 28 local pairs correctly. Proposition 9 nevertheless gives matrix loss four for this type against the faithful target. The separation therefore applies to a type capable of fitting every observed label. It remains a matrix-loss statement; the decoder's behavior on cross products is treated in Section 6.

### 4.2 Variation within the target type

**Example 11** (Correct type without target accuracy)**.** Take $X=I_t$ and $Y=\cos\theta I_t+\sin\theta J_t$, $0<\theta<\pi/2$. These define exact local representations with $M=-I_2$, zero alignment error, and six distinct local images, but

$$\mathcal{L}(S)=4(1-\sin\theta)>0. \tag{10}$$

At $\theta=\pi/6$ the loss is two; it tends continuously to zero as $\theta\to\pi/2$, without changing overlap type.

This family separates underdetermination from type mismatch. Even after the target's overlap type is fixed, the missing quaternion relation leaves a continuous choice. Hence improvement in matrix prediction need not involve a stratum change.

## 5 General Cyclic Overlap and the Evaluation Distribution

Let $C = \langle z \rangle$ have order $n$ and be central in both factors. For two overlap types, let $m_1(k), m_2(k)$ be the multiplicities of $\exp(2\pi \mathrm{i} k/n)$, and define

$$r_{\max} = \sum_{k=0}^{n-1} \min\{m_1(k), m_2(k)\}. \tag{11}$$

This is the largest number of eigenvalues that can be paired equally. Choose any distribution on base pairs $(a_0, b_0)$ and evaluate $(z^s a_0, b_0)$ with $s$ uniform on $\{0, \dots, n-1\}$. Base pairs outside $C$ give cross-region tests; uniform sampling on $A \times B$ also satisfies this condition.

**Proposition 12** (Cyclic-overlap separation lower bound)**.** *For the orbit sampling above and realizable types,*

$$c_{\rho_t}(\sigma_1; \mu) \geq c_{\mathrm{rel}}(\sigma_1, \sigma_t; \mu) \geq 2(d - r_{\max}), \tag{12}$$

*where* $\sigma_t = \sigma(\rho_t \circ \pi)$.

Character orthogonality gives a basis-free projector formula for the averaged cross term. Its nuclear norm (the sum of its singular values) is at most $r_{\max}$, even when the two overlap matrices do not commute. This proves the lower bound (Appendix E). Attainment requires the optimizing overlap actions and products to come from whole-factor representations consistently across all evaluated pairs. The bound is sharp for the relaxed quaternion targets of Proposition 9; fixed targets can have a larger separation, as the mixed type shows.

**Why sampling matters.** In the target coordinates of (17), take $X = \mathrm{diag}(1, -1)$ and $Y = XK_t$. Then $X^2 = Y^2 = I_2$, so the overlap type is wrong, but $XY = K_t$. This predictor has zero error on $(i, j)$ alone and average error four on the four-pair test. Disconnectedness permits a type mismatch; it does not imply positive loss for every observation distribution. The separation bound therefore requires the stated sampling condition.

**Scope of the overlap bound.** For trivial overlap, this particular separation bound is zero. Missing quotient relations can nevertheless cause errors. For infinite overlaps, path-connected representations need not be conjugate, so the finite-overlap alignment argument cannot be replaced by a connectedness test.

## 6 Reading the Stratum from Learned Operators

Theorem 7 is stated in terms of an operator-valued quasi-representation. Standard supervised grokking, the setup studied in the mechanistic-interpretability literature [16, 15, 3], does not directly produce such an operator: the network produces logits over the elements of $G$, with structure encoded in an embedding. The operator analog of group multiplication is obtained by fitting linear maps $L_g e_h \approx e_{gh}$. To convert this fit into the hypothesis of Theorem 7, both regions must be represented on one specified space $V$ with one inner product. Separate fits on different spans do not supply that shared space. For a region $H$, arrange the features in a matrix $E = [e_h]_{h \in H} \in \mathbb{C}^{d \times |H|}$ of full row rank. Let $P_g$ permute its columns by left multiplication in $H$.

**Proposition 13** (Measurable local defect)**.** *Suppose* $L_e = I_d$*,* $s = \sigma_{\min}(E) > 0$*, and*

$$r = \max_g \|L_g E - E P_g\|_{\mathrm{op}}, \qquad v = \max_g \|L_g^* L_g - I_d\|_{\mathrm{op}} < 1.$$

*Set* $q = 1 - \sqrt{1 - v}$ *and take the polar unitaries* $F(g) = L_g (L_g^* L_g)^{-1/2}$*. Then*

$$\max_g \|F(g) - L_g\|_{\mathrm{op}} \leq q, \qquad \max_{g,h} \|F(gh) - F(g)F(h)\|_{\mathrm{op}} \leq (2 + \sqrt{1+v})(r/s + q) =: b. \tag{13}$$

*If* $b < 1/10$*, an exact correction exists with* $\alpha \leq b + 120 b^2$.

The proof uses the group-label identity $(gh)k = g(hk)$, the smallest singular value, and a polar perturbation estimate (Appendix F). If each column residual is bounded by $\varepsilon'$, then $r \le \sqrt{|H|}\varepsilon'$. This relates the bound to per-example fitting errors without treating a small residual alone as a guarantee. The chosen metric matters: any whitening (a change of coordinates making the feature covariance the identity) must be specified and shared, and approximate unitarity must be checked in that metric. A directly computed defect can replace $b$ when it is smaller.

**Recovering a type.** For an exact representation of a finite group, the character determines its unitary conjugacy class [19]. For $C = \{e, z\}$, $\operatorname{tr} \rho(z) = d - 2m$; dividing this trace by $d$ gives the *normalized overlap character*. If $\|F(z) - \rho(z)\|_{\mathrm{op}} \le \alpha$, then

$$|\operatorname{Re}\operatorname{tr} F(z) - (d - 2m)| \le d\alpha. \tag{14}$$

When $d\alpha < 1$, rounding $(d - \operatorname{Re}\operatorname{tr} F(z))/2$ identifies $m$. Otherwise the bound may leave several types possible. Report the rank together with the normalized trace: zero trace at rank two and at rank six describes different multiplicities.

**Proposition 14** (Full-rank interpolation)**.** *Let $H$ have order $n$, and let its features $\{e_h\}_{h\in H}$ be linearly independent. On their span, the maps $L_g e_h = e_{gh}$ form the regular representation, are unitary for some inner product, and satisfy $\operatorname{tr} L_g = 0$ for $g \ne e$.*

This follows by transporting permutation matrices through the feature basis (Appendix F). It holds for random independent features without any training. On six independent observed quaternion features, the central element exchanges three pairs, again giving trace zero. This identifies the fitted involution, not the computation used by the network. The Fourier-weighted diagnostic in Appendix J also requires care: embedding energy need not measure downstream use.

**From matrix separation to classification.** A decoder that factors through $S(a, b)$ need not inherit its matrix loss. For example, the mixed diagonal operators in Section 4 have two distinct cross-product matrices, both different from all six local matrices. One decoder can label both cross products correctly; another can swap the cross labels while preserving every local label. Their operators and matrix losses agree.

A quantitative link requires additional information. For a supplied target codebook $T_g = \rho_t(g)$ with diameter $D = \max_{g,h} \|T_g - T_h\|_F > 0$, encode a classifier's prediction as $Z(a, b) = T_{\widehat{g}(a,b)}$. Then

$$\Pr_{\mu}[\widehat{g} \ne ab] \ge \frac{[\sqrt{c_{\rho_t}} - \sqrt{d}\,\delta - \kappa]_+^2}{D^2}. \tag{15}$$

Indeed $\mathcal{L}(Z) \le D^2 \Pr[\widehat{g} \ne ab]$; apply Theorem 7. Here $\kappa$ measures disagreement between actual coded outputs and operator products. The codebook supplies target information, including unseen label assignments, and estimating $\kappa$ requires access to the evaluated computation. Equation (15) is conditional, not a consequence of the embedding trace alone.

## 7 A Frozen-Representation Illustration

We evaluate two frozen four-dimensional representations under identical within-region supervision, training only the readout. Real four-dimensional operators supply real features to the readout; all matrix losses in this experiment are computed in that dimension. The experiment tests whether the two operator families support perfect local classification, whether the central probe distinguishes them, and how their matrix errors compare with classification performance. The representations are fixed throughout training, so their overlap types are controlled experimental conditions.

Arm F uses the realification of the faithful $Q_8$ target, its four-dimensional real-matrix form. Arm W uses

$$X = R \oplus I_2, \qquad Y = R \oplus \operatorname{diag}(-1, 1), \qquad R = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{16}$$

Both arms distinguish the six local elements and satisfy the local laws exactly. Their central matrices have normalized traces $-1$ and $0$. W's generators commute, so W is a representation of the amalgam that does not descend to the quaternion group.

In each arm, a width-32 ReLU readout maps the 16 entries of the frozen operator product to eight logits. Only this readout is trained, using the same 28 local pairs, three paired initialization seeds, and 2,500 full-batch AdamW updates. All runs and four checkpoints per run are retained. Appendix H specifies the optimizer, probe, numerical precision, and reproduction commands.

Table 1: Frozen four-dimensional representations. Accuracies are final values for every seed (three per arm). Matrix error is mean squared Frobenius distance on the four cross pairs, without dimension normalization. It is fixed before training. Probe rank refers to the span of flattened local matrices, not the native dimension four.

| Measurement | Faithful F | Wrong W |
|---|---|---|
| Within-region classification accuracy | $100\%$ | $100\%$ |
| Cross classification accuracy | $0\%$ | $0\%$ |
| Fixed-target matrix error | $0$ | $8$ |
| Native normalized overlap trace | $-1$ | $0$ |
| Fitted normalized central trace (rounded) | $-1$ | $0$ |
| Central-probe rank | $3$ | $4$ |
| Maximum central-fit residual | $7.61\times10^{-16}$ | $8.84\times10^{-16}$ |

The matrix gap eight has a direct proof: W's cross products are real symmetric and orthogonal, whereas the faithful cross targets are antisymmetric and orthogonal. Their Frobenius inner products vanish (Appendix H). This is an algebraic illustration, independent of the trained decoder. Both arms fit all local labels, but neither predicts the unseen labels $k, -k$ correctly in these runs. Those labels never occur as training targets. Therefore the classification outcome neither identifies a wrong overlap type nor turns the matrix gap into a classification-error gap.

The central fit acts on ranks three and four, respectively. Regional fit ranks are $(2, 2)$ for F and $(3, 4)$ for W. Their small residuals are useful consistency checks, but the separate regional fits do not meet Proposition 13's common-space hypothesis. Here the native local representations are supplied exactly. Frozen buffers remained unchanged, and all 24 saved checkpoints reproduce their saved logits exactly in the recorded environment.

## 8 Discussion and Conclusion

**What the theorem says and does not say.** The first half is constructive: corrected and aligned local representations give a consistent compositional predictor. The second half is the obstruction: that prediction can be systematically wrong, in a specific and quantifiable way. The quaternion example shows both sides of the distinction: the wrong types force a matrix gap, while the right type still leaves choices about the cross products. An informal walkthrough of this picture is given in Appendix I.

**Zero-shot setup vs. classical grokking.** In the strict zero-shot compositional split of this paper—training only on within-region pairs—the cross-region relations are unseen at training time. Classical grokking experiments can use random splits that include some cross-region samples. These are different sources of supervision. Missing cross-region labels alone do not establish a permanent plateau.

**Relation to mechanistic interpretability.** Mechanistic-interpretability accounts of grokking [15, 3] identify which features the grokked network has learned (irreps, Fourier modes) and when they emerge. The present framework is complementary: once the operators have been identified and checked, the certificate asks what they imply for unobserved compositions.

**Scope and limitations.** The theory assumes unitary local maps or controlled corrections to them. The finite-overlap alignment bound is quantitative in a specified common coordinate system; it is not invariant under independent changes of the modules' bases. A different type yields a positive lower bound only with appropriate evaluation coverage and sufficiently small perturbation and output-disagreement budgets. Sharpness of the cyclic bound depends on realizability across the evaluated pairs. Separate regional fits, uncalibrated Fourier energy, and a trace without rank or correction

uncertainty do not certify a network's prediction behavior. The six-run study fixes the representation by design and is too limited to characterize ordinary neural learning.

**Soft agreement and modular AI.** For modules that compose operators, *separately trained modules need not use the same internal basis; they need only be alignable—related by a conjugation—on the overlap.* The chosen unitary supplies that alignment. A small alignment cost controls the change to the local operators. Small $\ell$ does not on its own certify generalization to the target: two modules can both land in the same wrong stratum, giving small $\ell$ alongside a provable matrix-error gap. Local agreement with the target and fidelity of the output computation complete the prediction bound. Any additional relations of the task group must also be checked before the extension can be regarded as a representation of that group.

Determining when standard networks learn such computations, and whether their changes explain observed grokking transitions, requires further evidence about training. The present result supplies the conditional certificate and an exact worked example for that investigation.

## Funding and disclosures

**Funding and competing interests.** The author reports no third-party funding or support for this work and no relevant competing financial interests during the 36 months before submission.

**AI use statement.** We used AI tools to proofread the manuscript and improve its grammar and clarity, and to help write experimental code. AI also assisted with developing mathematical claims and proofs, designing experiments, and analysing their results. We used adversarial AI review to look for inconsistencies, missing assumptions, and errors in the mathematical arguments. Responsibility for the final text, mathematical claims, code, and results rests with the author.

## References


[1] E. Abbe, S. Bengio, A. Lotfi, and K. Rizk. Generalization on the unseen, logic reasoning and degree curriculum. In *Proc. ICML*, volume 202 of *PMLR*, pages 31–60, 2023.

[2] M. Burger, N. Ozawa, and A. Thom. On Ulam stability. *Israel J. Math.*, 193:109–129, 2013.

[3] B. Chughtai, L. Chan, and N. Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *Proc. ICML*, volume 202 of *PMLR*, pages 6243–6267, 2023.

[4] J. A. Fodor and Z. W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

[5] A. Gromov. Grokking modular arithmetic. *arXiv:2301.02679*, 2023.

[6] D. H. Hyers. On the stability of the linear functional equation. *Proc. Nat. Acad. Sci.*, 27(4):222–224, 1941.

[7] D. Kazhdan. On $\varepsilon$-representations. *Israel J. Math.*, 43:315–323, 1982.

[8] T. Kumar, B. Bordelon, S. J. Gershman, and C. Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *Proc. ICLR*, 2024.

[9] S. Lachapelle, D. Mahajan, I. Mitliagkas, and S. Lacoste-Julien. Additive decoders for latent variables identification and Cartesian-product extrapolation. In *Proc. NeurIPS*, 2023.

[10] B. Lake and M. Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proc. ICML*, volume 80 of *PMLR*, pages 2873–2882, 2018.

[11] S. Lippl and K. L. Stachenfeld. When does compositional structure yield compositional generalization? A kernel theory. In *Proc. ICLR*, 2025.

[12] Z. Liu, O. Kitouni, N. Nolte, E. J. Michaud, M. Tegmark, and M. Williams. Towards understanding grokking: An effective theory of representation learning. In *Proc. NeurIPS*, 2022.

[13] J. Lurie. *Higher Algebra*. Manuscript, 2017. Available at `https://www.math.ias.edu/~lurie/papers/HA.pdf`.

[14] G. L. Marchetti, D. Kunin, A. Myers, F. Acosta, and N. Miolane. Sequential group composition: A window into the mechanics of deep learning. In *Proc. ICML*, 2026. *arXiv:2602.03655*.

[15] N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt. Progress measures for grokking via mechanistic interpretability. In *Proc. ICLR*, 2023.

[16] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv:2201.02177*, 2022.

[17] P. Ram, T. Klinger, and A. G. Gray. What makes models compositional? A theoretical view: With supplement. *arXiv:2405.02350*, 2024.

[18] K. Sargsyan. Functorial neural architectures from higher inductive types. *arXiv:2603.16123*, 2026.

[19] J.-P. Serre. *Linear Representations of Finite Groups*. Springer, 1977.

[20] P. Wang and Z. Wang. Why neural network can discover symbolic structures with gradient-based training: An algebraic and geometric foundation for neurosymbolic reasoning. *arXiv:2506.21797*, 2025.

[21] S. M. Ulam. *A Collection of Mathematical Problems*. Interscience, 1960.

[22] E. R. van Kampen. On the connection between the fundamental groups of some related spaces. *Amer. J. Math.*, 55:261–267, 1933.

[23] T. Wiedemer, P. Mayilvahanan, M. Bethge, and W. Brendel. Compositional generalization from first principles. In *Proc. NeurIPS*, 2023.

# A Primer and the Space of Compatible Representations

**The quaternion task.** A group is a set with an associative multiplication, identity, and inverses. The quaternion group has elements $\{1, -1, i, -i, j, -j, k, -k\}$ and relations $i^2 = j^2 = k^2 = -1$, $ij = k$, and $ji = -k$. The local subgroups are $A = \{1, i, -1, -i\}$ and $B = \{1, j, -1, -j\}$, with intersection $C = \{1, -1\}$. Their within-region pair set has $16 + 16 - 4 = 28$ elements. It produces six labels; the cross pairs $(\pm i, \pm j)$ produce the two remaining labels $\pm k$.

A unitary representation assigns a unitary matrix to each group element and preserves multiplication. In complex dimension two a faithful quaternion representation, unique up to equivalence, is

$$I_t = \begin{pmatrix} \mathrm{i} & 0 \\ 0 & -\mathrm{i} \end{pmatrix}, \qquad J_t = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad K_t = I_t J_t. \tag{17}$$

All three square to $-I_2$, and $I_t J_t = -J_t I_t$. The overlap matrix of an arbitrary amalgam representation is a unitary involution. It can have both $+1$ and $-1$ eigenvalues; it need not be a scalar matrix.

**The universal property as a topological identification.** A representation of $A *_C B$ restricts to representations of $A$ and $B$ that agree on $C$. Conversely, the universal property extends every such compatible pair uniquely to the amalgam. Restriction and extension are therefore inverse maps between the two spaces.

Give representation spaces the topology of pointwise convergence: the matrix assigned to each group element must converge. Restriction is continuous because it simply keeps the matrices assigned to elements of $A$ and $B$. To see that extension is continuous, fix any element of the amalgam and express it as a finite word in the factors. Its matrix is a finite product of the corresponding local matrices, so it depends continuously on them. Both maps are thus continuous: the identification in Remark 1 is a homeomorphism.

For finite factors, all local matrices lie in the compact space $\mathrm{U}(d)^A \times \mathrm{U}(d)^B$. The representation laws and agreement on $C$ are closed matrix equations. Their solution set, the space of compatible pairs, is consequently closed and compact, even when the amalgam itself is infinite. This compactness will ensure that the continuous losses considered below attain their minima.

**Why types are components for finite overlap.** By complete reducibility, a unitary representation of finite $C$ is a direct sum of irreducible representations. If the irreducible indexed by $\lambda$ has dimension $d_\lambda$ and occurs $m_\lambda$ times, then $\sum_\lambda d_\lambda m_\lambda = d$. There are finitely many irreducibles and hence finitely many possible multiplicity vectors in dimension $d$.

Each vector specifies one unitary conjugacy orbit: changing the basis changes the matrices but not their multiplicities, and equal multiplicities give unitarily equivalent representations. The orbit of $\sigma$ is the image of $\mathrm{U}(d)$ under the continuous map $u \mapsto u\sigma u^*$. Since $\mathrm{U}(d)$ is compact and connected, so is each orbit. Each is also closed in the representation space, which is Hausdorff. Its complement is a finite union of other closed orbits, so the orbit is open as well. No connected subset can cross between these disjoint open and closed orbits; the orbits are therefore exactly the connected components, or overlap types.

For finite factors $A$ and $B$, the compatible pairs with a specified overlap type form a closed subset of the compact space above: restriction to $C$ is continuous and the chosen orbit is closed. If this subset is nonempty, any continuous finite weighted sum of squared matrix errors attains its minimum there. This proves the attainment assertion following Definition 5, without identifying its value.

**Normalization and stability.** The cited bound $\alpha_H \le \varepsilon_H + 120\varepsilon_H^2$ requires $F_H(e) = I_d$, unitary values, and a uniform defect below $1/10$ [2, Theorem 1.1]. If a unitary map instead has $\|F_H(e) - I_d\|_{\mathrm{op}} \le \nu_H$, normalize it by $F'_H(g) = F_H(e)^* F_H(g)$. Its change is at most $\nu_H$ and its multiplicative defect is at most $\varepsilon_H + \nu_H$: insert $F_H(e)^* F_H(g) F_H(h)$ between $F'_H(gh)$ and $F'_H(g)F'_H(h)$. Thus a correction for $F'_H$ has distance at most $\nu_H + (\varepsilon_H + \nu_H) + 120(\varepsilon_H + \nu_H)^2$ from the original map, when the normalized defect is below the threshold. The main text uses exact normalization to avoid this additional term.

The certificate does not require stability of the amalgam. It also applies to possibly infinite factors with finite overlap when exact corrections and uniform error bounds are supplied; replace factor

maxima by suprema. Amenable factors admit the same stability estimate. Finite factors are used in the main text so that all local-law and fitting checks are finite.

## B Proof of Finite-Overlap Alignment

*Proof of Lemma 6.* Define

$$T = \frac{1}{|C|} \sum_{c \in C} \sigma_2(c)\sigma_1(c)^*, \qquad u = T(T^*T)^{-1/2}.$$

Each summand differs from $I_d$ in operator norm by at most $\tau$, so $\|T - I_d\|_{\mathrm{op}} \le \tau < 1$ and $T$ is invertible. Reindexing the average by $c \mapsto hc$ gives

$$\sigma_2(h) T \sigma_1(h)^* = T.$$

Therefore $T\sigma_1(h) = \sigma_2(h)T$. Taking adjoints and using unitarity shows that $T^*T$ commutes with $\sigma_1(h)$. Its inverse square root also commutes with $\sigma_1(h)$ by functional calculus. Thus the polar unitary $u$ intertwines the representations and satisfies $u\sigma_1(h)u^* = \sigma_2(h)$.

For every vector $v$, the estimate on $T - I_d$ gives $(1-\tau)\|v\| \le \|Tv\| \le (1+\tau)\|v\|$. All singular values of $T$ lie in $[1-\tau, 1+\tau]$. Writing $T = u|T|$ yields $\|u - T\|_{\mathrm{op}} = \|I_d - |T|\|_{\mathrm{op}} \le \tau$. Finally $\|u - I_d\|_{\mathrm{op}} \le \|u - T\|_{\mathrm{op}} + \|T - I_d\|_{\mathrm{op}} \le 2\tau$. □

## C Proof of the Conditional Certificate

*Proof of Theorem 7.* Lemma 6 gives $u\widehat{\rho}_A|_C u^* = \widehat{\rho}_B|_C$. Hence $\rho_A = u\widehat{\rho}_A u^*$ and $\rho_B = \widehat{\rho}_B$ agree exactly on the overlap. The universal property gives a unique extension of this pair to $\widetilde{G}$. For a unitary matrix $M$,

$$\|uMu^* - M\|_{\mathrm{op}} = \|uM - Mu\|_{\mathrm{op}} \le \|(u - I_d)M\|_{\mathrm{op}} + \|M(u - I_d)\|_{\mathrm{op}} \le 2\ell.$$

Consequently $\|F_A(a) - \rho_A(a)\|_{\mathrm{op}} \le \alpha_A + 2\ell$ and $\|F_B(b) - \rho_B(b)\|_{\mathrm{op}} \le \alpha_B$. Decompose

$$S(a,b) - \widehat{y}(a,b) = [F_A(a) - \rho_A(a)]F_B(b) + \rho_A(a)[F_B(b) - \rho_B(b)].$$

Unitarity of the surrounding factors bounds its norm by $\delta = \alpha_A + \alpha_B + 2\ell$. Using $\ell \le 2\tau$ proves (3).

For a matrix-valued function $V$ write $\|V\|_{L^2_F(\mu)} = (\mathbb{E}_\mu \|V\|_F^2)^{1/2}$. The pointwise bound gives $\|S - \widehat{y}\|_{L^2_F(\mu)} \le \sqrt{d}\,\delta$. By the definition of fixed-target separation, $\|\widehat{y} - \rho_t(ab)\|_{L^2_F(\mu)} \ge \sqrt{c_{\rho_t}(\sigma_F; \mu)}$. The reverse triangle inequality therefore gives

$$\begin{aligned}\sqrt{\mathcal{L}(Z)} &\ge \|\widehat{y} - \rho_t(ab)\|_{L^2_F(\mu)} - \|S - \widehat{y}\|_{L^2_F(\mu)} - \|Z - S\|_{L^2_F(\mu)} \\ &\ge \sqrt{c_{\rho_t}(\sigma_F; \mu)} - \sqrt{d}\,\delta - \kappa.\end{aligned}$$

Taking the positive part and squaring proves (4).

For the target upper bound, expand instead

$$S(a,b) - \rho_t(ab) = [F_A(a) - \rho_t(a)]F_B(b) + \rho_t(a)[F_B(b) - \rho_t(b)].$$

Its operator norm is at most $\eta_A + \eta_B$, independently of overlap type. Passing to Frobenius and $L^2$ norms and adding $\|Z - S\|_{L^2_F(\mu)}$ proves (5). □

*Proof of Corollary 8.* For every $c \in C$, insert $F_A(c)$ and $F_B(c)$ between the two corrections. The triangle inequality gives $\tau \le \alpha_A + \zeta + \alpha_B$. Substitution in (3) yields the claimed bound. If the restrictions already agree, the identity aligns them. □

## D Exact Quaternion Calculations and Counterexamples

*Proof of Proposition 9.* In an amalgam representation, $M = X^2 = Y^2$ commutes with both $X$ and $Y$, and $M^2 = I_d$. Thus $\rho(-i) = MX$ and $\rho(-j) = MY$, proving (6).

*Fixed faithful target in dimension two.* If $m = 0$, $M = I_2$ and the average loss is

$$\tfrac{1}{2}\big(\|P - K_t\|_F^2 + \|P + K_t\|_F^2\big) = \|P\|_F^2 + \|K_t\|_F^2 = 4.$$

If $m = 1$, $M$ has two distinct eigenlines. Commutation with $M$ forces both $X$ and $Y$ to preserve them. On the positive eigenline, their eigenvalues are in $\{1, -1\}$; on the negative eigenline, they are in $\{\mathrm{i}, -\mathrm{i}\}$. Their products on both lines are real signs. Hence $P$ and $MP$ are Hermitian unitaries. The faithful $K_t$ obeys $K_t^2 = -I_2$ and $K_t^* = -K_t$. For a Hermitian $H$ and anti-Hermitian $K$, $\overline{\operatorname{tr}(HK)} = -\operatorname{tr}(KH) = -\operatorname{tr}(HK)$, so its real part is zero. Each cross error in (6) is therefore four. If $m = 2$, choosing $X = I_t$, $Y = J_t$ attains zero; nonnegativity shows this is the minimum.

*Relaxed target in any dimension.* Let the target be any amalgam representation with overlap $-I_d$ and cross product $Q$. With $\Pi$ the rank-$m$ projector onto the negative eigenspace of $M$, the average error is

$$\begin{aligned}\tfrac{1}{2}\big(\|P - Q\|_F^2 + \|MP + Q\|_F^2\big) &= 2d + \operatorname{Re}\operatorname{tr}(P^*(M - I_d)Q) \\ &= 2d - 2\operatorname{Re}\operatorname{tr}(\Pi Q P^*) \geq 2(d - m).\end{aligned}$$

The last inequality uses unitarity of $QP^*$. For attainment take

$$X = \operatorname{diag}(I_{d-m}, \mathrm{i}I_m), \quad Y = \operatorname{diag}(I_{d-m}, -\mathrm{i}I_m), \quad X_t = \mathrm{i}I_d, \quad Y_t = -\mathrm{i}I_d.$$

These satisfy the respective local fourth-power and shared-square relations, and give $P = Q = I_d$. The average is exactly $2(d - m)$. These relaxed targets need not descend to $Q_8$; that is why they cannot establish the mixed-type value for a fixed faithful target. □

*Calculation for Example 11.* The relations $I_t^2 = J_t^2 = -I_2$ and $I_t J_t = -J_t I_t$ imply $Y^2 = -I_2$ and $Y^* = -Y$. Thus $Y$ is unitary and the local relations hold exactly. For $0 < \theta < \pi/2$, the matrices $I_2, -I_2, \pm I_t, \pm Y$ are distinct. The cross product is $P = -\cos\theta I_2 + \sin\theta K_t$. Both predictor and target have the alternating pattern, so their average error is

$$\|P - K_t\|_F^2 = 2\cos^2\theta + 2(\sin\theta - 1)^2 = 4(1 - \sin\theta).$$

The middle equality uses the vanishing real Frobenius inner product of $I_2$ and $K_t$. □

*Proof of Proposition 10.* Since $\rho_H(e) = I_d$, the local pair $(e, h)$ has feature $\rho_H(h)$ and label $h$. If $\rho_H(h) = \rho_H(h')$, the decoder gives the same answer on both features. Correctness on all local pairs therefore implies $h = h'$, proving injectivity. For a nonidentity involution $c$, injectivity gives $\rho_H(c) \neq \rho_H(e) = I_d$. Its unitary matrix satisfies $\rho_H(c)^2 = I_d$, so its eigenvalues are $+1$ and $-1$. At least one must be $-1$.

The mixed quaternion witness $X = \operatorname{diag}(1, \mathrm{i})$, $Y = \operatorname{diag}(-1, \mathrm{i})$ has shared square $M = \operatorname{diag}(1, -1)$. Its six local images, displayed below, are distinct. Assigning each its corresponding group label defines a shared decoder on these images. Exact local multiplication gives the image of the correct label for every local pair, so this decoder classifies all 28 pairs correctly. Proposition 9 gives its fixed-target matrix error four. □

**A single-pair sampling counterexample.** With (17), $K_t = \left(\begin{smallmatrix} 0 & \mathrm{i} \\ \mathrm{i} & 0 \end{smallmatrix}\right)$. The matrix $X = \operatorname{diag}(1, -1)$ anticommutes with $K_t$. Hence $Y = XK_t$ obeys $Y^2 = -K_t^2 = I_2$ and is unitary. The shared square is $I_2$, but $XY = K_t$. The error at $(i, j)$ is zero; Proposition 9 gives four for the four-pair average.

**A product-decoder counterexample.** Set $X = \operatorname{diag}(1, \mathrm{i})$ and $Y = \operatorname{diag}(-1, \mathrm{i})$. The shared square is $M = \operatorname{diag}(1, -1)$. The six local matrices are

$$I_2, \quad M, \quad \operatorname{diag}(1, \mathrm{i}), \quad \operatorname{diag}(1, -\mathrm{i}), \quad \operatorname{diag}(-1, \mathrm{i}), \quad \operatorname{diag}(-1, -\mathrm{i}).$$

The two cross-product matrices are $P = \operatorname{diag}(-1, -1)$ and $MP = \operatorname{diag}(-1, 1)$, distinct from each other and from all local matrices. A function on these eight matrices can assign every local label correctly and assign either $(k, -k)$ or $(-k, k)$ to $(P, MP)$. Extending this finite function to a decoder proves the claim in Section 6. No change to the operators, local labels, or probe is involved.

## E Cyclic Separation and Nonuniform Coverage

*Proof of Proposition 12.* Fix two admissible representations and a base pair $(a_0, b_0)$. Write their overlap generators as $M_1, M_2$ and their base products as $P_1, P_2$. The orbit predictions are $M_1^s P_1$ and $M_2^s P_2$. Thus

$$\frac{1}{n}\sum_{s=0}^{n-1} \|M_1^s P_1 - M_2^s P_2\|_F^2 = 2d - 2\,\mathrm{Re}\,\mathrm{tr}(P_1^* D P_2), \qquad D = \frac{1}{n}\sum_s M_1^{-s} M_2^s.$$

For an $n$th root of unity $\omega$, let $\Pi_{j,\omega}$ be the spectral projector of $M_j$ at $\omega$. Expanding both powers and averaging characters gives

$$D = \sum_\omega \Pi_{1,\omega}\Pi_{2,\omega}.$$

This identity does not assume that $M_1$ and $M_2$ commute. Each product of projectors has operator norm at most one and rank at most $\min(m_1(\omega), m_2(\omega))$. Its nuclear norm is at most this rank. Consequently

$$\mathrm{Re}\,\mathrm{tr}(P_1^* D P_2) \le \|D\|_* \le \sum_\omega \min(m_1(\omega), m_2(\omega)) = r_{\max}.$$

This proves the lower bound for every base pair. Average over their distribution and take the joint infimum over the two representations. Restricting the second representation to $\rho_t \circ \pi$ can only increase that infimum. □

**Attainment and centrality.** Choosing matched diagonal overlap matrices and $P_1 = P_2 = I_d$ saturates the matrix inequality when those choices are allowed. It does not prove that such products are simultaneously realized by the factor representations for every evaluated base pair. The proof therefore establishes a lower bound, with equality only where witnesses are supplied. Left-orbit sampling itself does not use centrality. For central cyclic overlap, centrality additionally gives the two-sided pattern $\rho(z^s a_0)\rho(z^t b_0) = M^{s+t}P$. Uniform independent $s, t$ produce the same orbit average.

## F Operator Extraction, Characters, and Interpolation

*Proof of Proposition 13.* Let $R_g = L_g E - E P_g$. Because $P_g P_h = P_{gh}$,

$$(L_g L_h - L_{gh})E = L_g R_h + R_g P_h - R_{gh}.$$

The assumption on unitarity implies $\|L_g\|_{\mathrm{op}} \le \sqrt{1+v}$. Hence the right side has norm at most $(2+\sqrt{1+v})r$. Since $EE^\dagger = I_d$ and $\|E^\dagger\|_{\mathrm{op}} = 1/s$, it follows that

$$\|L_g L_h - L_{gh}\|_{\mathrm{op}} \le (2+\sqrt{1+v})r/s.$$

The singular values of $L_g$ lie in $[\sqrt{1-v}, \sqrt{1+v}]$. Its distance to its polar unitary is at most $q = 1 - \sqrt{1-v}$, since $\sqrt{1+v} - 1 \le q$. Also $F(e) = I_d$. Insert $L_{gh}$ and $L_g L_h$ between $F(gh)$ and $F(g)F(h)$. The product difference satisfies

$$\|L_g L_h - F(g)F(h)\|_{\mathrm{op}} \le \|L_g\|_{\mathrm{op}}\|L_h - F(h)\|_{\mathrm{op}} + \|L_g - F(g)\|_{\mathrm{op}}\|F(h)\|_{\mathrm{op}} \le (1+\sqrt{1+v})q.$$

Adding the remaining polar displacement $q$ proves (13). The finite-group stability theorem gives the final correction bound. □

**The column-residual version.** If every column of $R_g$ has Euclidean norm at most $\varepsilon'$, then $\|R_g\|_{\mathrm{op}} \le \|R_g\|_F \le \sqrt{|H|}\varepsilon'$. Since $q \le v$ and $\sqrt{1+v} < \sqrt{2}$,

$$b \le (2+\sqrt{2})\left(\frac{\sqrt{|H|}\varepsilon'}{s} + v\right).$$

This yields a residual-to-defect bound under explicit rank, metric, and normalization assumptions. The calculation uses only within-region labels. For two regions, however, the same $V$ and metric must be available; an estimate on one region's span cannot silently be applied to another span.

**Trace uncertainty.** For any matrix $M$, $|\operatorname{tr} M| \le d\|M\|_{\mathrm{op}}$. Applying this to $F(z) - \rho(z)$ proves (14). Consecutive admissible traces $d - 2m$ differ by two; uncertainty strictly below one therefore permits unique nearest-integer rounding for $m$. For a general finite overlap, multiplicities are recovered from the character inner products

$$m_\lambda = \frac{1}{|C|} \sum_{c \in C} \chi_\rho(c)\overline{\chi_\lambda(c)}.$$

If $\|F(c) - \rho(c)\|_{\mathrm{op}} \le \alpha$ uniformly, the error in this expression after replacing $\chi_\rho$ by $\operatorname{tr} F$ is at most $d\alpha|C|^{-1}\sum_c |\chi_\lambda(c)|$. These are uncertainty statements about an exact correction, not about neural feature attribution.

*Proof of Proposition 14.* Let $T : \mathbb{C}^n \to V$ send the standard basis vector $b_h$ to $e_h$. Independence makes $T$ invertible onto the feature span. The only maps satisfying the interpolation constraints are $L_g = TP_gT^{-1}$, so $L_gL_h = L_{gh}$. Equip $V$ with the inner product $\langle x, y\rangle_V = \langle T^{-1}x, T^{-1}y\rangle$. Then $L_g$ is unitary. Similarity preserves trace, and a nonidentity left translation on $H$ fixes no basis vectors, giving trace zero. For the six observed quaternion tokens, multiplication by $-1$ is a permutation of three disjoint transpositions. Its eigenvalues are three $+1$ and three $-1$, so the same interpolation argument gives zero trace. Other group multiplications can leave those six tokens; this does not construct a full $Q_8$ action on them. □

## G The Groupoid Formulation of Gluing

Van Kampen expresses global structure through compatible local data. For the higher-categorical background, Lurie's treatment of the Seifert–van Kampen theorem describes spaces as homotopy colimits [13, Appendix A.3]. The representation gluing needed here has the following explicit groupoid description.

Let $\mathcal{R}_d(H)$ be the groupoid whose objects are unitary representations of $H$ and whose morphisms are unitary intertwiners. The gluing groupoid has objects $(\rho_A, \rho_B, u)$ satisfying $u\rho_A(c)u^* = \rho_B(c)$ on $C$. A morphism to $(\rho'_A, \rho'_B, u')$ is a pair of intertwiners $(v_A, v_B)$ such that $v_Bu = u'v_A$.

Conjugate $\rho_A$ by $u$ and glue it to $\rho_B$ using the ordinary universal property. Send a morphism $(v_A, v_B)$ to $v_B$, which intertwines both aligned factors and therefore their extensions. This functor to $\mathcal{R}_d(A *_C B)$ is an equivalence: every global representation restricts to an object with $u = I_d$, and the compatibility equation uniquely determines $v_A$ from $v_B$. It is the groupoid (or homotopy-pullback) formulation of agreement up to a chosen basis change.

This construction concerns representations and their intertwiners. The mapping space from $BH$ to the topological space $B\mathrm{U}(d)$ instead classifies topological bundles, and need not describe flat representations with specified holonomy. For finite $C$, belonging to the same component guarantees that an intertwiner exists, while Lemma 6 supplies a quantitative choice when the restrictions are close in the current coordinates.

## H Experimental Protocol and Complete Results

**Experimental design.** The two arms share the training pairs, readout architecture, optimizer, and paired initialization seeds. Only the frozen operator family differs. The protocol fixes these choices before training and retains every specified run. Matrix errors and central actions are computed directly from the frozen operators; classification accuracy and cross-entropy are measured from the trained readouts.

**Operators and data.** Realify a complex matrix by replacing each entry $a + \mathrm{i}b$ with $\left(\begin{smallmatrix} a & -b \\ b & a \end{smallmatrix}\right)$. Arm F consists of the realified matrices of (17). For arm W, use $X, Y$ from (16) and the local table $(I_4, X^2, X, X^3, Y, Y^3)$ in order $(1, -1, i, -i, j, -j)$. Both tables contain six distinct orthogonal matrices and satisfy the local multiplication laws exactly. Their shared-square matrices are respectively $-I_4$ and $M = \operatorname{diag}(-1, -1, 1, 1)$. Training uses all 28 distinct local ordered pairs; evaluation uses the four cross pairs in the same order as (6). All eight output classes are present in the classifier, but only six occur as training targets. No cross outcomes select a seed or checkpoint.

**Algebraic loss eight.** For W, $P = XY = \mathrm{diag}(-1,-1,-1,1)$ and $MP$ are real symmetric orthogonal matrices. The realified target $Q = \rho_F(k)$ is antisymmetric and orthogonal because $Q^2 = -I_4$. For symmetric $H$ and antisymmetric $Q$, $\mathrm{tr}(H^TQ) = 0$. Consequently each of $\|P-Q\|_F^2$ and $\|MP+Q\|_F^2$ is $4+4=8$. The four-pair average is eight, while F has zero error. Simultaneous orthogonal conjugation of W preserves symmetry, so this value also holds after any such change of W's basis relative to the fixed target. These are deterministic algebraic values, not estimated means over training seeds.

**Readout and optimization.** For an input $(a,b)$, flatten the $4\times 4$ product of the frozen operators into 16 real entries. The readout is `Linear(16,32), ReLU, Linear(32,8)`, with biases. For each seed $0,1,2$, both arms use identical PyTorch-default initial readout weights. Train only readout parameters with full-batch cross-entropy for 2,500 AdamW updates, learning rate $0.001$, weight decay $1.0$, and betas $(0.9, 0.98)$. The rate is constant. Training uses float32, one CPU thread, and deterministic algorithms. There is no hyperparameter search, early stopping, or selection by cross accuracy. Save checkpoints before training and after 100, 500, and 2,500 updates, including optimizer states and raw logits. The frozen matrices are buffers, excluded from the optimizer, and are checked for exact invariance.

**Probe definition.** The observed feature for a local element $h$ is $e_h = \mathrm{vec}(\rho(h)) \in \mathbb{R}^{16}$. For the central probe, take an orthonormal SVD basis $Q$ for the span of all six features, retaining singular values above $10^{-10}$ times the largest. Fit $L = (Q^TE_{-h})(Q^TE_h)^\dagger$ and report $\mathrm{tr}(L)/\,\mathrm{rank}(E)$ and the maximum Euclidean column residual in ambient coordinates. This is left multiplication by the central matrix restricted to an invariant span. It gives rank three and normalized trace $-1$ for F, and rank four and normalized trace zero for W. Regional fits repeat the procedure separately on the four elements of each cyclic subgroup. Their ranks differ as reported in the main text; they are not presented as extraction on a shared four-dimensional feature space. Probe calculations use float64 on the supplied exact matrices and remain constant throughout readout training. The experiment therefore checks probe consistency, not recovery of changing learned operators.

Table 2: Every final run. CE denotes mean cross-entropy in nats. The matrix errors and overlap characters are deterministic and are reported in Table 1. No significance test or population-level success-rate estimate is inferred from three seeds.

| Arm | Seed | Train accuracy | Cross accuracy | Train CE | Cross CE |
|---|---|---|---|---|---|
| F | 0 | 1 | 0 | $7.663\times10^{-8}$ | 3.36626 |
| F | 1 | 1 | 0 | $9.366\times10^{-8}$ | 3.33867 |
| F | 2 | 1 | 0 | $1.362\times10^{-7}$ | 3.32353 |
| W | 0 | 1 | 0 | $3.661\times10^{-7}$ | 17.90069 |
| W | 1 | 1 | 0 | $5.450\times10^{-7}$ | 17.32113 |
| W | 2 | 1 | 0 | $9.196\times10^{-7}$ | 16.54741 |

**Reproduction and compute.** The accompanying code package contains `run.py`, `protocol.json`, environment versions, the complete result directory, and file hashes. It requires Python, NumPy, and PyTorch; all data are generated from the group matrices. Run

```
python run.py -out results_fresh
```

from the package directory. The script refuses to overwrite an existing output directory, runs all six cases, and reloads every checkpoint to verify logits. The saved environment and exact versions are included in `environment.json`. These are small CPU experiments with 808 trainable parameters per readout; no GPU or external dataset is required. Peak memory and wall-clock time were not recorded. Algebraic checks, frozen-buffer checks, identical paired initialization, and all 24 exact checkpoint replays passed. These numerical checks supplement the written proofs.

# I Informal Walkthrough

This appendix explains the quaternion example through two clocks. The precise statements are Theorem 7 and Proposition 9.

**Two clocks sharing midnight.** Picture $A = \langle i \rangle$ and $B = \langle j \rangle$ as two four-hour clocks. Each clock ticks $1 \to i \to -1 \to -i \to 1$ (or its $j$-analog). The within-region task is to learn each clock in isolation: a network that has memorized clock $A$ knows that $i \cdot i = -1$, that $i \cdot (-1) = -i$, and so on. The same for clock $B$. The clocks share 1 and $-1$; think of $-1$ as their shared "midnight." The cross-region task is to predict products that step from one clock to the other—$i \cdot j$, $i \cdot (-j)$, and so on. These products live in $Q_8$ but not in either clock alone.

**The hidden choice at midnight.** A local representation must respect $i^2 = -1$, but it need not send $-1$ to $-I$. In dimension two its matrix $M = \rho(-1)$ can be $+I$, $-I$, or an involution with one eigenvalue of each sign. The local multiplication laws allow all three possibilities. Learning distinct labels adds a constraint: if $M = +I$, the matrices for $i$ and $-i$ coincide, so a decoder receiving only the product matrix cannot distinguish those labels. The mixed type does allow perfect local classification, as the explicit construction in Section 4 shows. Thus exact local laws and correct local labels are different requirements.

The space of choices is $\mathrm{Hom}(C, \mathrm{U}(2))$, the representations of $C = \{1, -1\}$ in $2 \times 2$ unitary matrices. This space has three lumps—one for $\rho(-1) = +I$, one for $\rho(-1) = -I$, and one for the mixed case. These lumps are the *strata*. No continuous path within this space of exact representations carries one lump into another. The faithful target has $M = -I$; the other two types are wrong relative to that target.

**Why basis changes are allowed—but only inside a lump.** Two networks that have learned clock $A$ and clock $B$ separately will generally pick their own internal coordinates. Their matrices for $-1$ may not be literally equal. Within a lump they have the same eigenvalue multiplicities and can be aligned by a unitary basis change; between lumps they cannot. The groupoid formulation keeps track of this chosen alignment. Agreement between the two clocks is one question; agreement with the target is another.

**Why the wrong lump has a matrix-error gap.** Write $X = \rho(i)$, $Y = \rho(j)$, and $P = XY$. If $M = +I$, all four cross predictions are $P$, whereas the faithful target has the alternating pattern $(K_t, -K_t, -K_t, K_t)$. Constant cannot match alternating. The residual isn't a small misalignment; it's structural. Since $P$ and $K_t$ are unitary $2 \times 2$ matrices, the average squared Frobenius error is

$$\tfrac{1}{2}\big(\|P - K_t\|_F^2 + \|P + K_t\|_F^2\big) = \|P\|_F^2 + \|K_t\|_F^2 = 4.$$

The mixed type has the same fixed-target error for a different reason: its two eigenspaces force the predicted cross matrices to be Hermitian, while $K_t$ is anti-Hermitian. Their real Frobenius inner products vanish. This is the wrong type that can also fit all local labels. The number four describes a matrix separation; the decoder determines how those matrices become class labels.

**What the theorem says, informally.** Two halves.

**The constructive half.** If the operators approximately respect each clock's multiplication, stability supplies nearby exact clocks. Once the corrected clocks are aligned on their overlap, the universal property extends them uniquely to the amalgam $A *_C B$. The certificate bounds how much this correction and alignment change the products. To obtain a representation of $Q_8$ itself, the additional quaternion relation must also hold.

**The obstruction.** Two failure modes are possible. First, the clocks may have different overlap types, so no unitary alignment exists. Second, they may agree with each other but share a wrong type relative to the target. Then their exact extension has a positive gap on the specified cross pairs. Small approximation and output errors preserve a quantitative part of that gap. Conversely, agreement with the target's type alone does not fix the cross products: Section 4 gives a continuous family of errors within that type. Local target agreement supplies the separate upper bound in Theorem 7.

**What this picture does not capture.** The clocks describe exact representations and their controlled perturbations, not a trajectory of neural-network weights. They do not determine whether training will change type, how long that will take, or how a matrix gap will appear in classification loss. For larger central cyclic overlaps, character orthogonality replaces the constant-versus-alternating calculation and gives a lower bound under complete overlap-orbit sampling. Whether that bound is attained also depends on which factor representations and targets are available.

## J Connected Overlap and Fourier-Weighted Diagnostics

**Underdetermination with connected overlap.** In additive notation, take $G = \mathbb{Z}$, $A = 2\mathbb{Z}$, $B = 3\mathbb{Z}$, and $C = 6\mathbb{Z}$. Writing $a = 2$ and $b = 3$ gives $\widetilde{G} = \langle a, b \mid a^3 = b^2 \rangle$, the trefoil knot group. It differs from $G$: mapping $a$ to a three-cycle and $b$ to a transposition gives a surjection onto the nonabelian group $S_3$, whereas every quotient of $\mathbb{Z}$ is abelian. Thus the local data leave additional task relations unspecified. Nevertheless, $\mathrm{Hom}(C, \mathrm{U}(d))$ is homeomorphic to the connected space $\mathrm{U}(d)$: a representation is determined by its value at the generator 6. This example lies outside the finite-overlap certificate. Connectedness here does not imply conjugacy or remove the missing task relations.

**A Fourier-weighted diagnostic.** Let $H = \langle g \rangle$ have even order $n$ and let $c = g^{n/2}$. For features $e_j$ indexed by $g^j$, in a specified inner product, define

$$\widehat{e}_k = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} e^{-2\pi \mathrm{i} kj/n} e_j, \qquad w_k = \frac{\|\widehat{e}_k\|^2}{\sum_\ell \|\widehat{e}_\ell\|^2}, \qquad \chi_w = \sum_{k=0}^{n-1} w_k (-1)^k,$$

provided the denominator is nonzero. The sign is the value of the $k$th cyclic character at $c$. More generally, for supplied normalized nonnegative weights $w_\lambda$ on irreducibles and a central involution $c$, the corresponding weighted average is

$$\chi_w = \sum_\lambda w_\lambda \frac{\chi_\lambda(c)}{d_\lambda}, \qquad \frac{\chi_\lambda(c)}{d_\lambda} \in \{+1, -1\}.$$

The last assertion follows from Schur's lemma and $c^2 = e$. These formulas require no feature-rank collapse. Fourier energy weights describe the embedding; interpreting them as predictive usage requires a separate analysis of the readout. The weighted average can vary continuously, is not a discrete overlap type, and does not replace the character certificate. No claim about transition timing follows from it.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state a conditional operator certificate, the cyclic lower bound with orbit coverage, and the exact fixed-target quaternion losses. They explicitly distinguish matrix error from classification and learning dynamics.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 8 states the representation, correction, coordinate, sampling, and output-fidelity assumptions; Sections 4 and 6 give counterexamples. The experiment fixes representations and does not test their emergence.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The hypotheses are stated with the results. Appendices B–F give the proofs, including alignment, the certificate, quaternion equalities, the cyclic lower bound, and extraction. The stability theorem is cited with its normalization and threshold.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix H specifies the experimental protocol, all seeds, architecture, data, optimizer, probe, and stopping rule.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code and reproducibility package are provided as a separate supplementary upload with the paper. Appendix H gives the command and dependencies.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 7 gives the design and Appendix H specifies the full experimental protocol and per-seed results. Both frozen operator families and every run are specified.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: No inferential test or confidence interval is reported. Table 2 reports every final seed rather than only selected runs or an aggregate. The matrix gaps are exact algebraic quantities. No population inference or significance claim is made from three seeds.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix H specifies one-thread CPU execution, model size, updates, numerical precision, and software versions. Peak memory and wall-clock duration were not recorded.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work concerns finite mathematical tasks and a small synthetic experiment, with no personal data, human subjects, or high-risk deployed models. The experimental methods and AI assistance are disclosed.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational theory and a small algebraically generated illustration, without a direct application to consequential decisions. The limitations discourage treating an embedding probe as an unsupported guarantee of neural behavior.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released artifacts concern finite-group matrices and small readouts, not high-risk pretrained models or scraped datasets.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and respected?

    Answer: [No]

    Justification: No external dataset or pretrained model is used. The experiment uses Python, NumPy, and PyTorch, whose versions are recorded. A consolidated third-party license notice has not yet been assembled for a public release.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The experimental protocol, source, generated data, states, predictions, environment, and verification records are documented in Appendix H and the accompanying package.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subjects or crowdsourcing are involved.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects are involved.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The AI use statement describes writing and code assistance, mathematical development and adversarial review, and experimental design and analysis.