WHAT THE LOSS NEVER ASKS FOR: WHICH GROUP A RECURRENT NETWORK REPRESENTS, AND WHICH WORDS BREAK IT

Anonymous authors
Paper under double-blind review

ABSTRACT

Which finite-group state-tracking tasks a recurrent architecture can express is well characterised; which of several correct implementations training actually selects is not. We study this selection problem for LSTMs and input-switched linear RNNs trained on word problems over nine groups. In the linear models the state is a function of the running product if and only if the learned operators satisfy the group's defining relations on the states they reach; cross-entropy enforces them only as far as the margin requires at training lengths, and in ordinarily trained models they fail at order one. Identities that do hold, imposed by a penalty or by the architecture, present a group whose elements the state follows; training does not select the target group, which satisfies them all. Penalised on identities of the Klein four-group that also present an infinite group, models realise the infinite group: its word $(ab)^n$ breaks them within a thousand tokens while its finite-order words survive, and $(ab^3)^n$ breaks 1.96 times later (predicted 2). Penalised on identities of D_4, they realise its double cover, whose sign pattern appears in 44 of 48 models and in no control; one reflection per token, the delta rule's transition at $\beta = 2$, breaks first on $(xy)^n$, the travelling word of the infinite dihedral group. For isometric recurrences we bound how a relator defect accumulates and separate eigenplanes whose phase error drifts from those too small to reach the margin; snapping the two drifting planes to their ideal angles removes the break on the targeted word. A squared-norm variant of MatrixNet's relational penalty lengthens periodic-word horizons by one to three orders of magnitude without making any model exact; where it stalls, its floors are local minima of the penalty, derived in closed form for S_3. Instruments are calibrated on planted representations; we report all 19 pre-registered predictions, including the one that failed.

1 INTRODUCTION

Train an orthogonal input-switched linear recurrence on the Klein four-group $K_4 = \langle a, b \mid a^2, b^2, (ab)^2 \rangle$, with a penalty on three of its identities, a^4, a^2b^2 and ab^2a. K_4 satisfies them, and so does every representation of it; the trained models satisfy them to 3×10^{-5}. Yet the models do not represent K_4: a^2 acts as -1 on a large block of their states, so K_4's own relators fail at order one. The penalised identities also present an infinite group, $\langle a, b \mid a^4,\ b^2 = a^2 \rangle$, in which ab has infinite order while aba, bab and a^2b have order 4, and the models behave as that group predicts. They survive 20,000 random steps at the median and break within a thousand tokens on $(ab)^n$; $(aba)^n$, $(bab)^n$ and $(a^2b)^n$, which contain ab, stay unbroken in 42 of 42 models; and $(ab^3)^n$, which travels at half the speed, breaks 1.96 times later. The satisfied identities, not the target group, predicted the word that breaks the model. Random words would have missed it.

Architectures impose identities too. One reflection per token, the delta rule's transition (Schlag et al., 2021; Yang et al., 2024) at $\beta = 2$ with context-independent keys, satisfies $x^2 = e$ for every token by construction; with the rule's write term the map is an affine reflection and still an involution. On two reflection generators these identities present the infinite dihedral group D_∞. Trained on K_4, S_3 or D_4 written with two reflections, every model that breaks within 20,000 tokens breaks first on $(xy)^n$, at a median of 4,542 tokens, while random words survive.

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.

For a recurrent model there is a precise hypothesis about how a group is tracked: the state is the image of h_0 under a linear action of the element, $h(g) = T_g h_0$ (Foerster et al., 2017; Arjovsky et al., 2016; Henaff et al., 2016; Chughtai et al., 2023), and for non-solvable groups such as A_5 the word problem is NC^1-complete (Barrington, 1989; Merrill et al., 2024). The state depends only on the element exactly when the learned operators satisfy the group's *defining relations*, $r^3 = e$ or $(rs)^2 = e$, on the states they reach (Lemma 1). The data obey those relations, but cross-entropy asks the operators to satisfy them only as far as the margin needs at training lengths. That leaves open which implementation training selects: whether the relations hold at all and, when only some identities hold, *which group* the operators represent, a group $\widehat{G}$ that maps onto G without being it.

Two recent results frame the question. Chung et al. (2026) show that affine recurrences cannot correct error along the directions that separate states, so practical trackers learn finite horizons governed by accumulated error; Laird et al. (2024) learn generator matrices with a penalty on the defining relations and show that respecting them lets a model handle longer words. We ask what training does without such a penalty, and with identities that present a different group. The error's source is the relator defect on the reachable states (Proposition 1); identities that hold leave exactly the error they do not constrain, which predicts the word that breaks the model.

Contributions. (i) **Not exact within the tested budgets** (§3). The relators fail at order one on the occupied states of every ordinarily trained model we measured, on nine groups. (ii) **Which group training realises** (§4). When imposed identities hold, the state follows the elements of the group they present (Corollary 1); what is not forced is that training avoids the target group, which satisfies the same identities. It does: on K_4 and D_4 the models realise covers, whose travelling words break them first while their finite-order words survive far longer, the D_4 cover's sign pattern appears in 44 of 48 models and no control, and the breaking word moves with the omitted relator. (iii) **The mechanism of the horizon** (§5). For isometric recurrences a relator defect accumulates at most linearly, and along a periodic word it splits into eigenplanes whose phase error drifts and eigenplanes too small to reach the margin; snapping the two drifting planes removes the break in 90 of 90 models, and the certified horizon lies within a median factor of 2.9 of the observed break. (iv) **The relation penalty and its traps** (§6). A squared-norm variant of the penalty of Laird et al. (2024) drives the defect on the reachable states below 10^{-3} in most models on K_4, $\mathbb{Z}_6$, D_4 and $\mathbb{Z}_6 \times \mathbb{Z}_2$ and lengthens periodic horizons 48- to 1,700-fold, but no model is exact; on S_3, Q_8, A_4, S_4 and A_5 it stalls at floors set by local minima of the penalty, derived in closed form for S_3. Campaign F registered its predictions before evaluating them; Appendix F separates consistency checks from the outcomes the setup does not force, and reports the one that failed.

Scope. The propositions assume isometric linear recurrences; the unconstrained arm and the LSTM are reported empirically. Groups have order at most 60, and survival to 20,000 steps is censored evidence, not a guarantee.

2 Setup and instruments

Task and groups. For a finite group G with generating alphabet Σ, the model predicts $y_t = x_1 \cdots x_t$ at every timestep under cross-entropy over $|G|$ classes. The four original groups have short presentations, $\mathbb{Z}_6 = \langle a, a^{-1} \mid a^6, aa^{-1} \rangle$ (with a^{-1} a separate input token), $K_4 = \langle a, b \mid a^2, b^2, (ab)^2 \rangle$, $S_3 = \langle r, s \mid r^3, s^2, (rs)^2 \rangle$ and $D_4 = \langle r, s \mid r^4, s^2, (rs)^2 \rangle$; five further groups, $\mathbb{Z}_6 \times \mathbb{Z}_2$, Q_8, A_4, S_4 and the non-solvable A_5, are chosen to break the coincidences of the first four (Appendix A). Every relator is a positive word, and each presentation also presents G as a monoid (Appendix G), which §3 and §5 use.

Architectures. An **LSTM** (Hochreiter & Schmidhuber, 1997), whose state we take to be $[h; c]$, and an **input-switched linear RNN**, $h_t = A_{x_t} h_{t-1}$, with one $H \times H$ operator per generator and a learned h_0 (Foerster et al., 2017), so that a word $u = x_1 \cdots x_T$ acts as $A_u = A_{x_T} \cdots A_{x_1}$. The linear model comes in four arms: unconstrained (the *linear baseline*), exactly orthogonal, $A_x = \exp(W_x - W_x^\top) \in SO(H)$ at every step (the *orthogonal arm*; the exponential map of Lezcano-Casado & Martínez-Rubio (2019), not the Cayley transform, which cannot produce the eigenvalue -1, Appendix C), unconstrained with the relation penalty of §6, and a *reflection arm*,

$A_x = I - 2u_x u_x^\top / \|u_x\|^2$, one learned vector per token (the delta rule's transition at $\beta = 2$, with context-independent keys and no write term). Widths are 24 and 48 unless stated.

Training and evaluation. Adam (Kingma & Ba, 2015), learning rate 3×10^{-3}, batches of 64 from a fixed pool of 8,192 sequences, at most 3,000 steps, $L_{\text{train}} \in \{12, 24\}$, 24 seeds per condition. A run **converges** if its final training and fresh-sequence accuracies, and from Campaign B on its accuracy at every logged step of the final quarter, reach 0.99; only converged runs are measured unless stated. With two input tokens there are only $2^{12} = 4{,}096$ words of length 12, so at $L_{\text{train}} = 12$ the fresh-sequence check largely re-samples the pool; at 24 the overlap is negligible. Length generalisation is L_{break}, the first position beyond L_{train} at which accuracy falls below 0.90, evaluated in float64, orthogonal operators recomputed as $\exp(W_x - W_x^\top)$ in float64 from the stored parameters, to 20,000 steps. Campaign E adds *structured word families*, x^n for each generator, the alternating word $(x_1 x_2)^n$ and a biased random word (first letter with probability 0.9), with L_{break} for the deterministic families the first position at which accuracy over the preceding 100 positions falls below 0.90, while the random and biased families average over 32 sequences (Appendix E); lengths count tokens, so $(ab)^n$ has $2n$.

Six campaigns. An *arm* is a training recipe (architecture and penalty), a *cell* one arm on one group, width and training length, and a *placebo* a penalty on words other than the defining relators; a *campaign* is one batch of runs with rules frozen in advance. **A** is the original factorial (62,208 models); **B** retrains its core and adds placebos, further groups, patching and longer runs; **C**, **D** and **E** are follow-ups, E re-measuring the stored checkpoints of B–D under structured words, covering-group labellings and spectral analysis, and training the unconstrained-plus-penalty arm (384 models) and long controls (120); **F** registers 19 predictions before any evaluation and tests them on the stored checkpoints and on 720 new models (Appendix F).

The mean-span screen. For each element we collect the states reached by eight random words spelling it, centre them, and extract a basis of their span by SVD (rank tolerance 10^{-3}). Let η^2 be the share of the state's variance that lies between elements rather than within them, and η^2_{sub} the same on the span M of the per-element means. If the state is a function of the element, $\eta^2_{\text{sub}} = 1$; a model **passes** when $1 - \eta^2_{\text{sub}} < \tau$. The threshold is calibrated, not chosen: on planted representations with injected noise the instrument reads the code correctly down to $1 - \eta^2 = 5.0 \times 10^{-7}$ and misreads it at 2.8×10^{-6}, so $\tau = 10^{-6}$ was frozen between the two before any trained model was measured. A pass is a statement about M, not the whole state: the screen never failed an exact planted model, but it flags only 24 of 36 deliberately broken ones, so it can inflate positive counts (Appendix B.1), and it does not see a residual on the reachable span outside M (§6 shows the case). Structural claims below therefore rest on the relator defect on the reachable span, with the screen beside it.

Relator defect and the occupied span. Lemma 1 concerns the exact orbit span, which numerically is the whole state space in all 2,088 stored checkpoints we examined, though at most three-dimensional for the reflection arm (Appendix F). Every defect is therefore computed on the model's own parameters over the *occupied span* V, spanned by the sampled states above 10^{-3} of the largest singular value: we report $\|(A_w - I)\Pi_V\|_F / \|\Pi_V\|_F$ and a *visit-weighted* version, weighted by the state variance each direction carries, the one that tracks L_{break} (§5, Appendix B.3). Checking fitted operators for composition is vacuous at these widths (Appendix B.2). Finally, the *straight-line test*: a word u *travels* in a group if its element has infinite order there, so that u^n never returns. If the operators represent $\widehat{G}$, the structured families that travel in $\widehat{G}$ but cycle in G read $\widehat{G}$ off behaviour and expose failures that random words reach only diffusively (§4, §5).

3 Not exact within the tested budgets, and why the loss allows it

Two subspaces need to be kept apart: the orbit span $V = \text{span}\{A_w h_0 : w \in \Sigma^*\}$, which contains every state the model can produce and is invariant under the operators, and the mean span $M \subseteq V$, where η^2_{sub} is computed.

Lemma 1 (whole-state factorisation; proof in Appendix G)**.** *Suppose the relators $\mathcal{R}$ present G as a monoid. Then $A_u h_0$ depends only on the element that u spells, for every word u, if and only*

if $A_w|_V = I$ for every $w \in \mathcal{R}$. In that case $T_g := A_u|_V$ is well defined for any u spelling g, $T_{gh} = T_h T_g$, and $\rho(g) := T_{g^{-1}}$ is a representation of G on V.

Inserting or deleting a relator inside a word changes the state by $A_q(A_w - I)A_p h_0$, which vanishes when $A_w|_V = I$; conversely, appending a relator to any word leaves its element unchanged, so A_w fixes every orbit state and hence V. Any $1 - \eta^2_{\text{sub}} > 0$, be it 0.6 or 4×10^{-4}, means that some relator fails to act as the identity where the states live; a pass, computed on M, does not establish the converse on V.

Not exact, measured on the operators. On the models' own operators the relators fail at order one where the states live. The visit-weighted defect, which weights each direction of V by the variance the states carry there and is the one that tracks L_{break} (§5), is 0.25 at the median in the orthogonal arm. The unweighted restricted defect, dominated by barely visited directions, is at least 0.97 in all 756 ordinarily trained models whose operators we measured, on nine groups (Figure 7, Appendix B.3). No screen threshold is involved. The mean-span screen agrees: none of the 57,920 converged models of Campaign A's factorial passes it, nor any of 600 orthogonal models across nine groups (Figure 3 in Appendix D).

Why training does not repair it. Cross-entropy rewards the correct logit for being the largest and keeps widening the margin inside that cone, but nothing in it requires $A_w|_V = I$: states of different words spelling the same element may differ in any direction that keeps them in the cone. Since within-element spread costs margin and gradient descent on separable cross-entropy grows margins (Soudry et al., 2018; Lyu & Li, 2020; Morwani et al., 2024), the relations may be approached in the limit; within our budgets they are not. The readout sees the remainder, yet at training lengths it is inert (Appendix E).

Longer training. The orthogonal arm trained to 30,000 steps at a constant rate passes at none of 30 checkpoints (it has no scheduled long run; Appendix D). Under a cosine schedule with weight decay the unconstrained linear residual falls from 0.711 at 3,000 steps to 0.053 at 60,000, still 4.7 orders of magnitude above τ, while the LSTM's *rises* from 0.352–0.376 to 0.96 at 100,000 steps with train and held-out accuracy at 1.0000 (Figure 6); these are the arms a grokking reading would ask for (Power et al., 2022; Nanda et al., 2023; Morwani et al., 2024).

4 Which group did it learn?

Identities imposed on a recurrence hold exactly when its architecture imposes them and approximately when a penalty does, and together they present a group $\widehat{G}$ that maps onto G.

Corollary 1 (proof in Appendix G)**.** *The four redundant-word lists below, the subsets of relators and $x^2 = y^2 = e$ each make both generators invertible, so they present $\widehat{G}$ as a monoid. If they act as the identity on V, the state is a function of the element of $\widehat{G}$; if they hold to ε on isometries, a word of finite order in $\widehat{G}$ drifts by at most its area times $\varepsilon\|h_0\|$ per period.*

The survival of finite-order words is therefore a consistency check, which the tests below measure. What the setup does not force is which group training realises, since G itself satisfies every imposed identity. The prediction is a separation of time scales: a travelling word drifts at the angle error of the unconstrained directions and breaks the model first, provided its drift is visible to the readout (Proposition 2), while words of finite order in $\widehat{G}$ drift only at the far smaller ε, and random words travel only diffusively. It could have failed: a realised finite quotient would make the travelling word cycle, generic drift would break every word containing the travelling factor, and tail-plane drift would survive the surgery of §5. Campaign F registered these tests before its evaluation scripts were written; Appendix F separates consistency checks from the outcomes the setup does not force.

Redundant words. The first placebo penalises other identities of G, the first ones outside $\mathcal{R}$ in shortlex order (Appendix E). Table 1 gives the groups they present, found by coset enumeration (Todd & Coxeter, 1936) or by inspection; the S_3 list presents S_3 itself, so that row is an alternative presentation rather than a placebo, and the $\mathbb{Z}_6$ row amounts to omitting a^6, because $a^{-1}a = e$ and $aa^{-1} = e$ are equivalent for invertible operators. Under random words every row survives 20,000

Table 1: Redundant-word penalties and their registered tests (orthogonal arm, $\beta = 0.1$, 48 models per row). $\widehat{G}$: the group the penalised words present. Met: share of models whose penalised words hold to 10^{-3} on V (on $\mathbb{Z}_6$ the other 34 sit in a local minimum of the penalty). Every row survives 20,000 random steps at the median; worst: the earliest-breaking structured family, median over converged models.

G	penalised words	$\widehat{G}$	met	worst family	registered test (Campaign F)
S_3	$(sr)^2,\ s^4,\ r^3s^2$	S_3	0%	none breaks	—
D_4	$(sr)^2,\ s^4,\ r^4s^2$	SD_{16}	25%	none breaks	SD_{16} sign pattern on a common block in 44/48; controls 0/48
K_4	$a^4,\ a^2b^2,\ ab^2a$	infinite	88%	$(ab)^n$: 954	$(aba)^n$, $(bab)^n$, $(a^2b)^n$ survive in 42/42; $(ab^3)^n$ breaks 1.96× later (predicted 2)
$\mathbb{Z}_6$	$a^3a^{-3},\ a^{-1}a$	$\mathbb{Z}$	29%	a^n: 660	$(a^ka^{-j})^n$ breaks at $(k+j)/\lvert k-j\rvert$ times a^n, within 4%

steps at the median. Under structured words the two infinite cases break on their travelling words, K_4 at a median of 954 tokens on $(ab)^n$ over its 42 converged models (822 over all 48) and $\mathbb{Z}_6$ at 660 on a^n, while a^n and b^n, of order 4 in K_4's $\widehat{G}$, survive, as do all families of the finite cases (Table 1; every family in Appendix E).

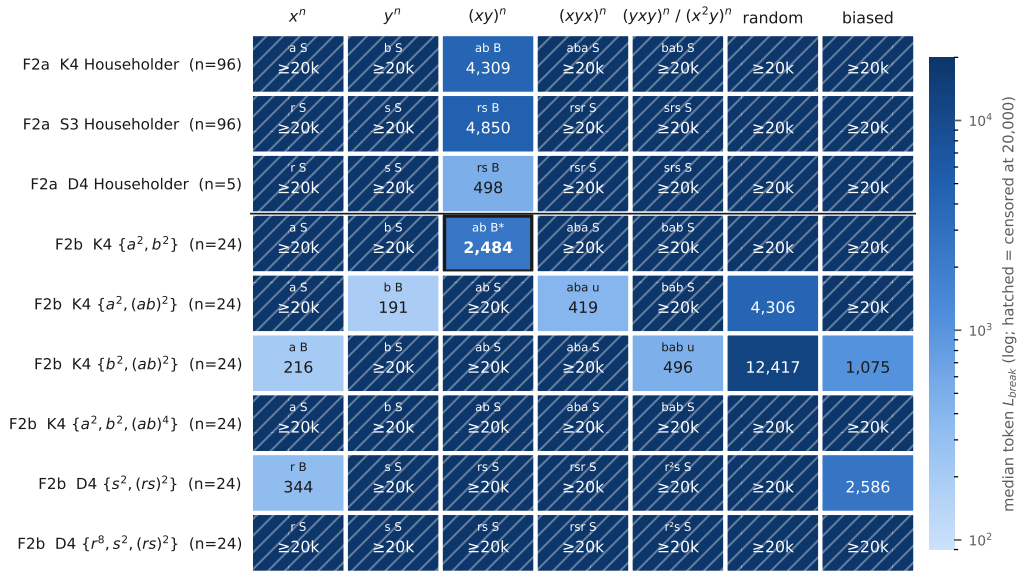
Discriminating words. Drift along the one unpenalised direction would break every word built from it; representing $\widehat{G}$ is a stronger claim, and the registered families separate the two. In K_4's $\widehat{G}$, aba, bab and a^2b contain ab yet have order 4, and $ab^3 = a^2 \cdot ab$ travels like ab at half the speed per token. In all 42 K_4 models that meet their penalised words, $(aba)^n$, $(bab)^n$ and $(a^2b)^n$ never break within 20,000 tokens, and $(ab^3)^n$ breaks 1.96 times later than $(ab)^n$ in the median model (IQR 1.94–1.98; predicted 2). On $\mathbb{Z}_6$, $(a^ka^{-j})^n$ survives when $k = j$ and otherwise breaks at $(k + j)/|k - j|$ times the a^n break, to within 4%.

The D_4 cover. In SD_{16}, r^4 and s^2 map to the central element and $(rs)^2$ to the identity. In 44 of the 48 D_4 models, all the converged ones, r^4 and s^2 act as -1 on one common block invariant under A_r and A_s, on which $(rs)^2$ has no eigenvalue -1. The joint pattern appears in none of the 48 un-penalised controls and none of 1,000 random orthogonal pairs, although single relators with a -1 eigenvalue are common in both. Labelling the reachable states by the elements of SD_{16} fits them better than labelling by D_{16}, Q_{16} or D_4 in 48 of 48 models.

Moving the break. Penalising a subset of the defining relators moves the travelling word (Figure 1). On K_4, $\{a^2, b^2\}$ presents D_∞ and leaves $(ab)^n$ to travel, $\{a^2, (ab)^2\}$ leaves b^n, and $\{b^2, (ab)^2\}$ leaves a^n; on D_4, $\{s^2, (rs)^2\}$ leaves r^n; $\{a^2, b^2, (ab)^4\}$ and $\{r^8, s^2, (rs)^2\}$ present finite groups. In five of the six sets every registered family behaved as predicted. The exception, $\{a^2, b^2\}$, broke on $(ab)^n$ and on no other family, but at a median of 2,484 tokens, above the registered cutoff of one tenth of the random-word median (2,000).

Identities imposed by the architecture. A transition that is a single reflection, the delta rule's at $\beta = 2$ (Schlag et al., 2021; Yang et al., 2024) with context-independent keys, satisfies $x^2 = e$ for every token by construction, so on two reflection generators the architecture presents D_∞, in which xy has infinite order and x, y, xyx and yxy are involutions. Trained on K_4, and on S_3 and D_4 written with two reflections, 183 of the 197 converged models break before 20,000 tokens, every one of them first on $(xy)^n$, at a median of 4,542 tokens, while random words survive; within K_4 and within S_3 the break falls with the angle error of A_yA_x relative to $2\pi/m$ at Spearman -0.999. The involution words cannot break once the few states of their orbits are decoded correctly. A learned $\beta \in (0, 2)$ (Grazzi et al., 2025) imposes no such identity. Only 5 of 96 D_4 models converged, for reasons we have not established.

What this identifies, and what it does not. The identification is strongest on K_4, approximate on D_4 (the sign pattern is exact but only 25% of the models meet the cover's words to 10^{-3}), and behavioural on $\mathbb{Z}_6$, where 34 of 48 models sit in a local minimum of the penalty in which the operator of $a^{-1}a$ rotates one plane by $114.09°$ (Appendix F). Exact satisfaction would still identify only an image of $\widehat{G}$, possibly a proper quotient, which a finite horizon cannot separate from $\widehat{G}$. LSTMs, outside Lemma 1, show run-dependent structured failures (Appendix E).



B/S: registered to break/survive; u: unregistered; *: registered prediction not met (median < 20,000 but not < random/10).
(x, y) = (a, b) on K4, (r, s) on S3 and D4. Random and biased words: Campaign E estimator, 32 sequences. F2a rows: converged models only.

Figure 1: The broken word follows the imposed identities (Campaign F, registered). Median token L_{break} by word family; hatched cells are censored at 20,000. Top: single-reflection transitions, which impose $x^2 = y^2 = e$ and so present D_∞ (converged models; D_4: 5). Bottom: penalties on subsets of the defining relators, each presenting the group named. In the reflection rows for S_3 and D_4 the cell labels r and s stand for the two reflection generators $x = s$ and $y = sr$. The outlined cell is the one registered prediction that failed. Row labels start with the registered prediction they test (Table 14); Householder is the reflection arm.

5 Drift versus oscillation

Throughout, V is the orbit span and the operators A_x are isometries of V, which is exact for the orthogonal arm. The unconstrained arm is outside these hypotheses and is reported empirically (Appendix E). The readout is $\ell(h) = Wh + b$ and the margin of a state at class c is $\mu(h, c) = \ell_c(h) - \max_{k \neq c} \ell_k(h)$; a positive margin means a correct read. Two facts suffice: an isometry does not amplify a perturbation, and a logit perturbation of norm δ moves the margin by at most 2δ.

Linear accumulation. Chung et al. (2026) prove that affine recurrences cannot correct error along state-separating directions. For group tracking, that error is a relator defect. Write $\mathcal{R}$ for the defining relators and $\varepsilon := \max_{w \in \mathcal{R}} \|(A_w - I)|_V\|$ for the worst relator defect on V in operator norm. A normalised Frobenius defect d on the same span brackets it, $d \leq \varepsilon \leq \sqrt{\dim V}\, d$, but the defect of §2, taken on the occupied span, does not bound it on omitted directions. An *elementary move* inserts or deletes a relator as a subword; since $\mathcal{R}$ presents G as a monoid, two words spell the same element iff a chain of moves connects them, and the least length of such a chain is their *area*. Fix a normal form $n(g)$ for every g, with $n(e)$ empty, and let c_G be the largest area between $n(g)x$ and $n(gx)$ over the finitely many pairs (g, x).

Proposition 1 (accumulation bound; proof in Appendix G)**.** *If u and u' spell the same element and are at area N, then $\|A_u h_0 - A_{u'} h_0\| \leq N\varepsilon\|h_0\|$. Consequently the state of any word of length T lies within $c_G T\, \varepsilon\|h_0\|$ of the state of the normal form of its element, and if every normal form is read with margin at least $\mu > 0$, every word of length $T < \mu / \big(2\|W\| c_G \varepsilon \|h_0\|\big)$ is read correctly.*

This is a sufficient guarantee, not a prediction of where the model breaks. Empirically, over the 271 uncensored models of Campaign C, L_{break} correlates with the visit-weighted defect (Spearman -0.86) about as strongly as with $1 - \eta^2_{\text{sub}}$ (-0.80; Appendix B.3); the two part ways on models whose residual stays bounded (Remark 2). For the un-penalised orthogonal controls and the penalised $\mathbb{Z}_6$ models the product $a \times L_{\text{break}}$, with $a = \sqrt{1 - \eta^2_{\text{sub}}}$ the residual amplitude, is 4.7, 6.7 and 4.6 at the median across three decades of a (Figure 2a), a constant that is empirical and not the proposition's. One Campaign-A cell (S_3, $\beta = 10^{-3}$) passes no model, yet reaches a median L_{break} of 1,719, against 102 for its control.

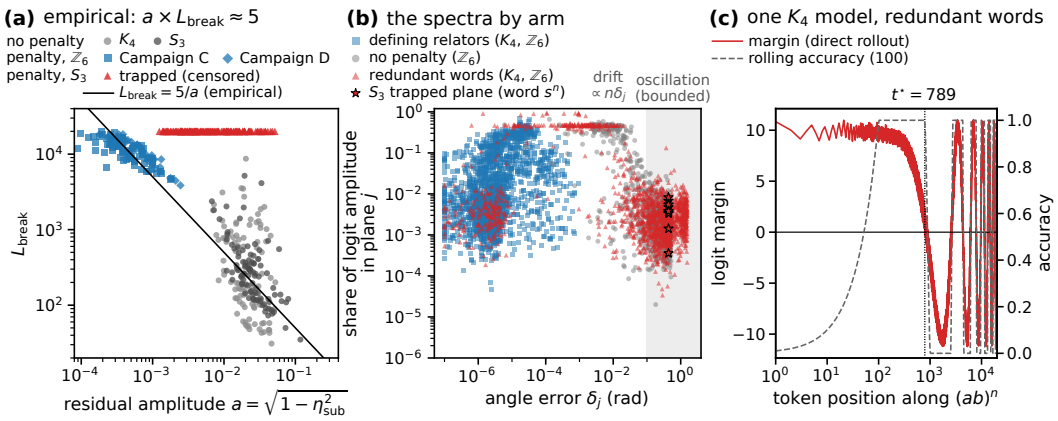


Figure 2: Drift versus oscillation. (a) L_{break} against the residual amplitude $a = \sqrt{1-\eta_{\text{sub}}^2}$ for the uncensored models of the un-penalised orthogonal controls and the penalised $\mathbb{Z}_6$ cells (Campaigns C and D), with the empirical line $5/a$; a is not the operator-norm defect ε of Proposition 1. The trapped S_3 models (red triangles) are censored at 20,000, where that line would put them near 1,200. **(b)** Per eigenplane of the travelling element's operator, angle error against share of the logit amplitude, by arm, with the S_3 trapped plane (stars) from eight Campaign C checkpoints. Every plane of an isometry is bounded; small angle errors drift over long horizons, and a plane is harmless only if its amplitude is small against the margin. **(c)** One redundant-word K_4 model under $(ab)^n$: the logit margin from direct rollout (red) and the rolling accuracy (grey), against token position; the first negative margin, at token 789, precedes the windowed break at 822.

Periodic words. A relator defect need not accumulate: it may rotate a component of the state around a circle it never leaves. Take a word u whose element g_u has order m in G and let $M := A_u|_V$. Decompose V into real orthogonal M-invariant blocks P_j, rotation planes and lines with eigenvalue ± 1, and write $h_0 = \sum_j h_j$ with $h_j \in P_j$. On each plane replace the rotation angle φ_j by a nearest multiple of $2\pi/m$, at circular distance δ_j, the *angle error*; keep a $+1$ line, keep a -1 line when m is even and replace it by $+1$ when m is odd, with $\delta_j = \pi$. This defines an isometry M^0 with $(M^0)^m = I$; let μ^0 be the least margin along the m-periodic ideal orbit $(M^0)^n h_0$, each state judged at the class of g_u^n, and let $g_j := \max_{c \neq k} \|\Pi_{P_j}(w_c - w_k)\|$ be the readout's logit-difference gain on block j, with w_c the row of W for class c.

Proposition 2 (drift versus oscillation; proof in Appendix G)**.** *For every* $n \geq 0$,

$$\left\|M^n h_0 - (M^0)^n h_0\right\| \;\leq\; \sum_j \|h_j\| \min(n\delta_j,\, 2),$$

$$\mu\big(M^n h_0,\, g_u^n\big) \;\geq\; \mu^0 - \sum_j g_j \|h_j\| \min(n\delta_j,\, 2).$$

For any partition of the blocks into $\mathcal{O}$ *and* $\mathcal{D}$*, let* $B = \mu^0 - 2\sum_{j\in\mathcal{O}} g_j\|h_j\|$ *and* $C = \sum_{j\in\mathcal{D}} g_j\|h_j\|\delta_j$*. If* $B > 0$*, the endpoint of* u^n *is read correctly for every* $n < B/C$*, and for every* n *if* $C = 0$*.*

Blocks with small δ_j *drift*; blocks whose readout-visible amplitude cannot exhaust the margin at any phase *oscillate*. A large angle error alone does not make a block harmless.

Remark 1 (reconstruction and surgery). For a block $u = x_1 \cdots x_q$ the logits at token $nq + r$ are $W A_{x_1 \cdots x_r} M^n h_0 + b$, $0 \leq r < q$, an explicit trigonometric sum over the eigenplanes of M; replaying them reproduces the break of all 90 redundant-word K_4 and $\mathbb{Z}_6$ models, which checks the decomposition, not a mechanism. The registered intervention tests the mechanism (Appendix F): in these models two dominant planes carry 92% of the readout-visible amplitude at angle errors near 10^{-3}, and snapping only them to their ideal angles removes the break on the targeted word in 90 of 90 models, out to 10^6 tokens, while snapping every other plane, or two random tail planes, moves the median break by 0.3% and 0.4%. With the dominant planes as $\mathcal{D}$, $B > 0$ in 87 of 90 models, and the certified horizon B/C lies below the observed break by a median factor of only 2.9 (minimum 1.6).

Remark 2 (the trapped plane). The trapped S_3 configuration (§6) contains a jointly invariant plane on which the relators act as rotations with angle errors of order one. In eight trapped checkpoints it carries 2–7% of the state's energy and 14–28% of the readout's weight, and trapped models survive every word family to 20,000: a bounded, low-amplitude component, although certifying it at every length would need $B > 0$ from phase-wise margins.

6 ASKING FOR THE RELATIONS

Laird et al. (2024) add a penalty on the defining relations, a sum of unsquared norms $\|A_r - I\|$, to learn group representations, and find that respecting the relations lets models handle longer words. We use its squared variant to test the account of §§3–5:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{task}} \;+\; \beta \sum_{w \in \mathcal{R}} \|A_w - I\|_F^2, \tag{1}$$

with $\beta \in \{10^{-3}, 10^{-1}, 1\}$, on all four groups, both widths and both lengths, 24 seeds each, at 3,000 steps under a constant learning rate.

Small defects, at no cost in accuracy. On the orthogonal arm **650 of 1,152 models pass** the screen (659 on retraining), against none in 192 matched controls, and every model converges. Measured on the reachable span the counts hold except on S_3: at $\beta \in \{0.1, 1\}$ a restricted defect below 10^{-3} holds in 192 of 192 K_4, 170 of 192 D_4, 182 of 192 $\mathbb{Z}_6$ and 13 of 192 S_3 models, against 192, 171, 184 and 35 screen passes; the 22 S_3 models that pass only the screen each carry one trapped plane (below) that lies in V but outside the mean span (Table 17).

Longer horizons, but finite ones. Every one of the 384 penalised models matched to Campaign C's controls survives 20,000 random steps, against control medians of 60 to 840 (Appendix H). Periodic words can be followed further: their break, computed from the eigendecomposition and validated against direct rollout (Appendix F), comes before 10^7 tokens for every one of the 1,770 evaluated periodic words of orthogonal models that survived 20,000. Models trained on the defining relators break on periodic words at medians of 7,582 to 266,676 tokens, 48 to 1,700 times their un-penalised controls: the penalty buys two to three orders of magnitude, not exactness.

Orthogonality is not needed on the four original groups. Campaign E runs Eq. (1) on the unconstrained linear RNN (384 models, same seeds, pool and optimiser). At $\beta = 1$ all four groups pass the screen (0.58 to 1.00) and survive 20,000 random and structured steps at the median; at $\beta = 0.1$ D_4, K_4 and S_3 do the same and $\mathbb{Z}_6$ does not; at $\beta = 10^{-3}$ nothing passes. The operators are not orthogonal, and the propositions of §5 are not invoked for this arm. A 3,000-step cosine decay seemed to undo the penalty (no model passed at $\beta = 0.1$); the decay halves the integrated learning rate, and with the integral matched 40 of 96 models pass the screen and 65 of 96 have a defect below 10^{-3} on V, against 0 and 23 (Appendix I). This arm has a trap of its own: 18 of 48 S_3 models at $\beta \geq 0.1$ sit in a real local minimum of the penalty, one direction of V being an eigenvector of A_r with eigenvalue -0.496 and nearly one of A_s with eigenvalue 1.084; they break within a hundred steps on s^n, yet 15 of them pass the screen, the direction lying outside the mean span (Appendix E).

Where asking is not enough: traps. On the five further groups the penalty makes the abelian $\mathbb{Z}_6 \times \mathbb{Z}_2$ pass (83%) and fails on Q_8, A_4, S_4 and A_5 (Appendix J). It is not inert there, reducing the defect by a factor of 1.6 to 5, and every penalised model still survives to 20,000 random steps, but it stops three to four orders of magnitude short of a representation, and more width makes it worse, not better. The defect reaches its floor within 500 steps and stays there over the observed training, and the floors come in whole-number multiples of a group-specific constant: 32 of 32 trapped A_4 models and 35 of 36 S_4 models within 2%, among Campaign D's 36 penalised models per group (Q_8: 19 of 36, against about 20% expected by chance at its levels; A_5's floors, sums over seven constants, are no test). Each constant is a local minimum of the penalty on a small invariant block, recovered from random starts with no measured input (Appendix J). For S_3 the trap is solved exactly: on a 2-plane invariant under both generators, on which A_s rotates by θ and A_r by α, the three relators compete for two angles, and the penalty $P = 4[3 - \cos 3\alpha - \cos 2\theta - \cos(2\alpha + 2\theta)]$ has a critical point at $\alpha = \pi - 2\theta$, $\sin^2 2\theta = 7/12$, with value $12 - 20\sqrt{15}/9$ and Hessian $\frac{8\sqrt{15}}{3}\left(\begin{smallmatrix} 4 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, whose eigenvalues $\frac{8\sqrt{15}}{3}(3 \pm \sqrt{2})$, 16.378 and 45.590, make it a strict local minimum of the restricted penalty, an obstruction of the landscape, not of expressivity. It is critical for the squared penalty only (the unsquared one has gradient 1.53 there and a non-smooth minimum of its own, of value $\sqrt{6}$, on the same plane), and trained checkpoints sit near it in a slow region of the joint objective (Appendix E). The relation target of Appendix K is then $c\sqrt{k}$ with $c^2 = 6 - 10\sqrt{15}/9$ and k the number of complex eigenvalues of A_s, two per trapped plane; the measured value agrees to a relative

2×10^{-7} and the measured angles to 0.04σ. At an explicit $SO(6)$ construction the full penalty's Hessian is positive definite off the gauge orbit (Appendix K), and because the point is an exact critical point of the penalty, $\nabla(\beta P) = 0$ for every β: no β moves the plateau, and neither tested schedule does (Appendix I). The trapped plane is a low-amplitude component (Remark 2): trapped S_3 models survive every word family to 20,000 while their $1 - \eta^2_{\mathrm{sub}}$ sits at 10^{-5} to 10^{-3}.

7 Related work

State tracking and length generalisation. That a recurrent network's states drift under long inputs unless the dynamics are attracting is the oldest result on RNNs and automata (Zeng et al., 1993; Casey, 1996; Omlin & Giles, 1996), and recurrent and Transformer models fail to generalise in length on regular tasks under random-string evaluation (Delétang et al., 2023). Chung et al. (2026) sharpen it for affine recurrences, which cannot correct error along state-separating directions and so learn finite horizons governed by accumulated error; their result bounds horizons by accumulated error, and §5 identifies that error with relator defects in the isometric case, predicts from the presentation which words expose it, and tests the prediction by intervention. Buitrago & Gu (2025) tie recurrent length generalisation to the coverage of attainable states; our failures lie along a direction that periodic words drive linearly and random words only diffusively (Appendix E). Expressivity results say what an architecture can represent (Barrington, 1989; Merrill et al., 2024; Grazzi et al., 2025; Terzić et al., 2025; Shakerinava et al., 2026; Siems et al., 2025; Sarrof et al., 2024; Fan et al., 2024); we ask what training selects, and find a cover. Transformers learn shortcuts that fail out of length (Liu et al., 2023b; Li et al., 2025), rare inputs expose failures that typical data hide (Liu et al., 2023a), and automata extracted from trained recurrent networks can mislead (Kolen, 1994) while words that break them can be searched for (Weiss et al., 2018) or held out by design (Lee, 2026); ours are chosen by the presentation, which also names the identity a model violates and the group it satisfies instead.

Group operations and structural constraints. Feedforward networks on group composition have been reverse-engineered as multiplying irreducible representations (Chughtai et al., 2023; Nanda et al., 2023; Stander et al., 2024; Wu et al., 2025; Morwani et al., 2024; Zhong et al., 2023); those claims concern which irreps a single step uses, not the functional dependence of a recurrent state, and none asks which *group* the representation is of. Group representations have been learned under homomorphism constraints (Quessard et al., 2020; Keurti et al., 2023), and MatrixNet (Laird et al., 2024) penalises the defining relations and reports longer-word generalisation: Eq. (1) is a squared-norm variant of it, which we use as a probe, not as a contribution. Operators with small relator defects are close to an exact representation (Kazhdan, 1982; Gowers & Hatami, 2017), which would give a word-independent certificate and repair; we have not computed it. Like soft-equivariance regularisers (Finzi et al., 2021; Wang et al., 2022) it puts group structure in the objective. §2 joins work on illusory interpretability conclusions (Bolukbasi et al., 2021; Makelov et al., 2024; Hewitt & Liang, 2019; Adebayo et al., 2018).

8 Limitations and conclusion

Propositions 1 and 2 assume isometric recurrences; the unconstrained arm and the LSTM, whose structured failures vary between runs, are outside them. The reflection arm converged on 5 of 96 D_4 models, and one registered prediction failed on its threshold. There is no long run of the orthogonal arm under a decaying schedule. We test chosen word families, not a search over words. Campaign F's timing rests on our git history and left two choices unspecified, and Campaign E's registration carries deviations (Appendix E). The task is synthetic, the models small, and random words stay censored at 20,000.

The identities a recurrence satisfies, imposed or not, present a group, and the model breaks first on that group's words of infinite order. Which identities hold, and which group training realises, are worth asking wherever a model tracks structure it was never told.

AI USE STATEMENT

In this work, we used generative AI tools for literature search, research ideation, orchestrating timed code runs, data analysis, and drafting of the manuscript. The authors set the research questions and evaluation criteria, approved each pre-registration before it was frozen, decided scope and claims, and reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Code, per-model records, the stored checkpoints behind every figure and table, and the frozen run specifications will be released with the paper. Thresholds (τ_{rank}, τ_{int}, τ), the group manifests and the hypothesis set were frozen with separate hashes before the data they govern. Campaign F's 19 predictions were hashed before any evaluation and applied as written (Appendix F); Campaign E's structured-word rule was hashed before its test ran, and Appendix E gives its text, its hash and the five ways its evaluation departed from it. The validation gate of Appendix B.1 reproduces in 3.6 s on CPU; every follow-up report is generated by a script that prints each number beside the query that produced it, and every guard was verified by breaking its input on purpose (28/28). Proofs are in Appendix G and the S_3 derivation in Appendix K.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1120–1128. PMLR, 2016.

David A. Barrington. Bounded-width polynomial-size branching programs recognize exactly those languages in NC^1. *Journal of Computer and System Sciences*, 38(1):150–164, 1989.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for BERT, 2021.

Ricardo Buitrago and Albert Gu. Understanding and improving length generalization in recurrent models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5856–5874. PMLR, 2025.

Mike Casey. The dynamics of discrete-time computation, with application to recurrent neural networks and finite state machine extraction. *Neural Computation*, 8(6):1135–1178, 1996.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6243–6267. PMLR, 2023.

Jiwan Chung, Heechan Choi, and Seon Joo Kim. Rethinking state tracking in recurrent models through error control dynamics, 2026.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. Neural networks and the Chomsky hierarchy. In *The Eleventh International Conference on Learning Representations*, 2023.

Ting-Han Fan, Ta-Chung Chi, and Alexander Rudnicky. Advancing regular language reasoning in linear recurrent neural networks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (Short Papers)*, 2024.

Marc Finzi, Gregory Benton, and Andrew Gordon Wilson. Residual pathway priors for soft equivariance constraints. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Jakob N. Foerster, Justin Gilmer, Jascha Sohl-Dickstein, Jan Chorowski, and David Sussillo. Input switched affine networks: An RNN architecture designed for interpretability. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1136–1145. PMLR, 2017.

W. T. Gowers and O. Hatami. Inverse and stability theorems for approximate representations of finite groups. *Sbornik: Mathematics*, 208(12):1784–1817, 2017. doi: 10.1070/SM8872.

Riccardo Grazzi, Julien Siems, Arber Zela, Jörg K. H. Franke, Frank Hutter, and Massimiliano Pontil. Unlocking state-tracking in linear RNNs through negative eigenvalues. In *The Thirteenth International Conference on Learning Representations*, 2025.

Kyle Helfrich, Devin Willmott, and Qiang Ye. Orthogonal recurrent neural networks with scaled Cayley transform. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1969–1978. PMLR, 2018.

Mikael Henaff, Arthur Szlam, and Yann LeCun. Recurrent orthogonal networks and long-memory tasks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, 2016.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 2733–2743. Association for Computational Linguistics, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

David Kazhdan. On ε-representations. *Israel Journal of Mathematics*, 43(4):315–323, 1982.

Hamza Keurti, Hsiao-Ru Pan, Michel Besserve, Benjamin F. Grewe, and Bernhard Schölkopf. Homomorphism AutoEncoder – learning group structured representations from observed transitions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 16190–16215. PMLR, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

John F. Kolen. Fool's gold: Extracting finite state machines from recurrent network dynamics. In *Advances in Neural Information Processing Systems*, volume 6, 1994.

Lucas Laird, Circe Hsu, Asilata Bapat, and Robin Walters. MatrixNet: Learning over symmetry groups using learned group representations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1021.

Jeonghoon Lee. A held-out transition-pair falsifier for long-horizon non-Abelian state tracking, 2026.

Mario Lezcano-Casado and David Martínez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3794–3803. PMLR, 2019.

Belinda Z. Li, Zifan Carl Guo, and Jacob Andreas. (how) do language models track state? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34429–34452. PMLR, 2025.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Exposing attention glitches with flip-flop language modeling. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *The Eleventh International Conference on Learning Representations*, 2023b.

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

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *The Twelfth International Conference on Learning Representations*, 2024.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35492–35506. PMLR, 2024.

Depen Morwani, Benjamin L. Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham Kakade. Feature emergence via margin maximization: Case studies in algebraic tasks. In *International Conference on Learning Representations (ICLR)*, 2024.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023.

Christian W. Omlin and C. Lee Giles. Constructing deterministic finite-state automata in recurrent neural networks. *Journal of the ACM*, 43(6):937–972, 1996.

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

Robin Quessard, Thomas D. Barrett, and William R. Clements. Learning group structure and disentangled representations of dynamical environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yash Sarrof, Yana Veitsman, and Michael Hahn. The expressive capacity of state space models: A formal language perspective. In *Advances in Neural Information Processing Systems*, volume 37, pp. 41202–41241, 2024.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 2021.

Mehran Shakerinava, Behnoush Khavari, Siamak Ravanbakhsh, and Sarath Chandar. The expressive limits of diagonal SSMs for state-tracking. In *The Fourteenth International Conference on Learning Representations*, 2026.

Julien Siems, Timur Carstensen, Arber Zela, Frank Hutter, Massimiliano Pontil, and Riccardo Grazzi. DeltaProduct: Improving state-tracking in linear RNNs via Householder products. In *Advances in Neural Information Processing Systems*, 2025.

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.

Dashiell Stander, Qinan Yu, Honglu Fan, and Stella Biderman. Grokking group multiplication with cosets. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46441–46467. PMLR, 2024.

Aleksandar Terzić, Michael Hersche, Giacomo Camposampiero, Thomas Hofmann, Abu Sebastian, and Abbas Rahimi. On the expressiveness and length generalization of selective state-space models on regular languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20876–20884, 2025. doi: 10.1609/aaai.v39i19.34301.

J. A. Todd and H. S. M. Coxeter. A practical method for enumerating cosets of a finite abstract group. *Proceedings of the Edinburgh Mathematical Society*, 5(1):26–34, 1936.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 2022.

Gail Weiss, Yoav Goldberg, and Eran Yahav. Extracting automata from recurrent neural networks using queries and counterexamples. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5247–5256. PMLR, 2018.

Wilson Wu, Louis Jaburi, Jacob Drori, and Jason Gross. Towards a unified and verified understanding of group-operation networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Zheng Zeng, Rodney M. Goodman, and Padhraic Smyth. Learning finite state machines with self-clustering recurrent networks. *Neural Computation*, 5(6):976–990, 1993.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27223–27250, 2023.

A THE NINE GROUPS

Table 2: The four original groups. $d_{\min}$ is the minimal dimension of a faithful real representation.

G	$\lvert G\rvert$	structure	generators	relators	$\lvert G^{\mathrm{ab}}\rvert$	$d_{\min}$
$\mathbb{Z}_6$	6	abelian	$\{a, a^{-1}\}$	$a^6,\ aa^{-1}$	6	2
K_4	4	$\mathbb{Z}_2 \times \mathbb{Z}_2$	$\{a, b\}$	$a^2,\ b^2,\ (ab)^2$	4	2
S_3	6	$\mathbb{Z}_3 \rtimes \mathbb{Z}_2$	$\{r, s\}$	$r^3,\ s^2,\ (rs)^2$	2	2
D_4	8	$\mathbb{Z}_4 \rtimes \mathbb{Z}_2$	$\{r, s\}$	$r^4,\ s^2,\ (rs)^2$	4	2

The four original groups (Table 2) are all solvable, have at most eight elements and have $d_{\min} = 2$. The five further groups are chosen in pairs that differ in one property at a time: the abelian $\mathbb{Z}_6 \times \mathbb{Z}_2$ (order 12, $d_{\min} = 3$); $Q_8 = \langle a, b \mid a^4, a^2b^2, abab^3\rangle$, which has the same order and character degrees as D_4 but a quaternionic irrep, so that $d_{\min} = 4$, and no generator of order 2; and A_4, S_4 and the non-solvable A_5, presented as $\langle a, b \mid a^2, b^n, (ab)^3\rangle$ with $n = 3, 4, 5$. Every relator is a positive word, checked against an independent reference implementation. Table 22 (Appendix J) lists all nine with their orders and $d_{\min}$.

B THE INSTRUMENTS

B.1 THE MEAN-SPAN SCREEN: EXTRACTION, VALIDATION AND CALIBRATION

Extraction. For each element we collect the states reached by eight distinct random words spelling it, with lengths spread over $[1, L_{\mathrm{train}}]$; states are centred before the SVD, and the constant direction is counted separately, which is the honest decomposition of an affine action. The rank tolerance $\tau_{\mathrm{rank}} = 10^{-3}$ is relative to the top centred singular value; sweeping it over two decades at widths 24 and 48 changes the pass rate by half a percentage point. The *code* of a planted representation is its decomposition into irreducible representations; the instrument reads it from the characters $\chi(g) = \operatorname{tr} A_g$ of operators fitted to the per-element means, and "reading the code correctly" means recovering the planted multiplicities. This is the only use of fitted operators in the paper; they are never taken as evidence that a model respects the group (Appendix B.2).

Validation. The instrument was validated before it was used, on planted fixtures with known representations: 192 configurations (4 groups $\times$ 16 codes $\times$ 3 widths, 40 seeds, 7,680 models), every one recovered exactly. Of 36 deliberately broken models it flags 24 and assigns a confident code to twelve, so it can inflate positive counts; no exact planted model was rejected. A pass does not certify the whole state: variation outside the span of the per-element means is not seen. Its guards fired on 13 of 13 broken inputs. The screen tests whether the state is a function of the element on that span, not which function: a perturbation applied identically to every instance of an element leaves the within-element variance at zero and is, correctly, not flagged. The relator-defect measure of §3 was validated separately: a planted representation gives $\approx 10^{-15}$, and one with a generator rescaled by 1.10 gives 0.249.

Calibration. Exact representations of the four original groups are planted in a larger state, their generators perturbed as $\rho(x)\exp(\varepsilon M_x)$ with M_x skew, and states generated along real words, so the perturbation accumulates as it would in a model (248 models per ε; Table 3). The threshold 10^{-6} falls between $\varepsilon = 10^{-4}$ and 3×10^{-4}, that is, at a relator defect of about 2×10^{-3}, and the orthogonal arm's typical 5×10^{-4} corresponds to a defect of a few hundredths. The relationship is monotone on average but loose, since at a fixed defect the induced $1 - \eta^2_{\mathrm{sub}}$ varies over several orders of magnitude, which is why the paper reports η^2_{sub} and the defect side by side.

B.2 THE COMPOSITION CHECK CANNOT FAIL

Fitting one operator per generator to the per-element means and checking that the fitted operators compose is vacuous once the state has more than $|G|$ dimensions, since any permutation of affinely independent points is realised exactly by an affine map, and by a linear one once a constant coordinate is appended (§2). Measured, for $|G| = 6$, composition residuals of 6.9×10^{-1} at $D = 3$ and 4.5×10^{-1} at $D = 4$ collapse to 1.1×10^{-14} at $D = 5 = |G| - 1$, and every width in this paper is

Table 3: Planted perturbation ε, the relator defect it produces, and the induced $1-\eta^2_{\mathrm{sub}}$ (median, minimum and maximum over 248 models).

ε	relator defect	$1-\eta^2_{\mathrm{sub}}$ median	min	max
0	6.3×10^{-16}	0	0	0
1×10^{-6}	8.5×10^{-6}	2.0×10^{-11}	5.2×10^{-14}	1.1×10^{-8}
1×10^{-5}	8.3×10^{-5}	2.3×10^{-9}	2.6×10^{-11}	9.9×10^{-7}
1×10^{-4}	8.7×10^{-4}	2.2×10^{-7}	7.8×10^{-10}	3.4×10^{-6}
3×10^{-4}	2.5×10^{-3}	2.1×10^{-6}	3.1×10^{-11}	6.6×10^{-4}
1×10^{-3}	8.4×10^{-3}	2.4×10^{-5}	2.0×10^{-7}	1.4×10^{-2}
3×10^{-3}	2.5×10^{-2}	2.8×10^{-4}	1.6×10^{-7}	2.6×10^{-2}
1×10^{-2}	8.7×10^{-2}	3.4×10^{-3}	3.8×10^{-5}	6.9×10^{-2}
3×10^{-2}	2.6×10^{-1}	3.0×10^{-2}	3.6×10^{-4}	8.7×10^{-1}
1×10^{-1}	9.0×10^{-1}	2.6×10^{-1}	10.0×10^{-4}	9.3×10^{-1}
3×10^{-1}	2.2×10^{0}	8.1×10^{-1}	2.0×10^{-2}	9.5×10^{-1}

in the latter regime. Three demonstrations. *Gaussian noise decodes as the regular representation*: random vectors are in general position, so the check returns the regular representation for all four groups and all 32 seeds and passes every guard including faithfulness. *A deliberately wrong Cayley table returns a clean code*, with $\varepsilon_{\mathrm{hom}} \approx 10^{-15}$, because the operators were fitted to whatever transitions we supplied; in the language of Hewitt & Liang (2019) this is a control task on which the check has no selectivity. *The diagnostic improves as the model gets worse*: moderate corruption of a planted code is flagged, but heavier corruption restores full rank and the composition error falls by twelve orders of magnitude. This is why the screen is variance-based and why every relator defect in the paper is computed on the model's own parameters. The argument concerns the per-element *means*; a composition residual over all sampled transitions is over-determined and is not vacuous.

B.3 The restricted and visit-weighted relator defect

Estimator. Collect states X (64 words per element); U = left singular vectors of centred X together with the mean direction, rank at 10^{-3} of the top singular value; $\Pi_V = UU^\top$; $\mathrm{defect}_V(w) = \|(A_w - I)\Pi_V\|_F / \|\Pi_V\|_F$ with A_w the product of the model's own A_x. The visit-weighted defect replaces Π_V by $C^{1/2}$, with C the second-moment matrix of the collected states: $\|(A_w - I)C^{1/2}\|_F / \|C^{1/2}\|_F$. Here V is the numerical span of the sampled states; finite sampling and rank truncation do not certify invariance or exhaust the exact orbit span. No least-squares operator fit appears anywhere: an estimator that uses operators *fitted* to the per-element means reads 10^{-15} for every model, including linear-baseline models with $\eta^2_{\mathrm{sub}} = 0.33$, which the contrapositive of §3 forbids (the counting argument of §2 in another guise). Fixtures: an exact representation planted on a proper subspace beside a rotation that violates the relators gives 4.6×10^{-16} (restricted) and 6×10^{-16} (visit-weighted) against a parameter-space defect of 8.3; generators perturbed at $\varepsilon = 10^{-2}$ give 3.4×10^{-2} and 3.5×10^{-2}; a real linear-baseline checkpoint with $\eta^2_{\mathrm{sub}} = 0.44$ gives 9.4 and 21.

Which defect predicts L_{break}. Spearman correlations with L_{break} over the 271 uncensored models of Campaign C: visit-weighted defect -0.86, $1-\eta^2_{\mathrm{sub}}$ -0.80, unweighted restricted defect -0.49, parameter-space defect -0.55; within the orthogonal arm ($n = 192$, none censored): -0.63, -0.47, $+0.41$, $+0.23$. The unweighted defect has the opposite sign within the orthogonal arm: it weights every direction of V equally although most are barely visited (Table 4), and weighting by visitation restores the expected sign in this population. The correlations are computed on uncensored models, which are mostly un-penalised controls.

C Why the exponential map and not the Cayley transform

For every evaluation the operators are recomputed from the stored skew parameters as $\exp(W_x - W_x^\top)$ in float64, not cast from float32 matrices; Campaign F measures their orthogonality residual below 10^{-12} and their eigenvalue moduli within 10^{-14} of 1.

An order-2 generator forces -1 eigenvalues: $s^2 = e$ means $A_s^2 = I$, so every eigenvalue of A_s is ±1, and faithfulness requires at least one -1. The plain Cayley transform $(I+S)^{-1}(I-S)$ of a skew-symmetric S cannot produce one, since its eigenvalues are $(1-i\lambda)/(1+i\lambda)$ for the imaginary

Table 4: Rank of the reachable span at tolerance 10^{-3} against the number of directions holding 99% and 99.9% of the state-cloud variance, and the participation ratio (medians over the stored checkpoints of Campaigns B–D; "24/48" pools both widths, so the median $\dim V$ of the orthogonal arm, which fills the width, is the median of the two widths).

arm	n	width	$\dim V$	$d_{99\%}$	$d_{99.9\%}$	participation ratio
linear baseline	384	24/48	28	9	17	2.9
LSTM baseline	384	24/48	66	18	34	4.8
orthogonal	192	24/48	36	14	30	3.5
orth. + penalty	1152	24/48	8	5	5	4.5
placebos (i), (iv), (v)	480	24	11	5	7	4.0

eigenvalues $i\lambda$ of S, which reach -1 only as $\lambda \to \infty$; scoRNN's scaled variant (Helfrich et al., 2018) adds a ± 1 diagonal for exactly this reason. Measured, $\exp(W - W^\top)$ at $\theta = \pi$ gives eigenvalues $\{-1, -1, 1, 1, 1, 1\}$ with $E^2 = I$ exactly, whereas the plain Cayley transform's minimum eigenvalue real part on random skew input is -0.855. A plain-Cayley search would therefore have reported zero representations for S_3, K_4 and D_4 for a reason having nothing to do with the hypothesis under test. The orthogonality error of the exponential parametrisation is 1.3×10^{-6} in fp32, with $\det = +1$.

D NOT EXACT WITHIN THE TESTED BUDGETS: FACTORIAL, REPLICATION, SCALE AND LONG TRAINING

Campaign A trains 62,208 models (11.3 GPU-hours on one consumer GPU) in 81 arms that cross weight decay, initialisation scale, a soft orthogonality penalty on the recurrent kernel and a length curriculum at three levels each, the lowest level of each being off and the orthogonality penalty topping out at 10^{-1}, on 4 groups, 2 training lengths, 2 widths and 2 architectures with 24 seeds each; 57,920 converge (93.1%), and only those are measured (Table 5). The LSTM on $\mathbb{Z}_6$ trains poorly at both widths and is reported separately, never pooled. The only knob that matters is the soft orthogonality penalty, and its effect has opposite signs on the two architectures: on the linear RNN it improves η^2_{sub} from about 0.35 to above 0.99; on the LSTM it makes things worse, because there it constrains W_{hh}, which is only a proxy for an effective transition that the gates make input-dependent. Even at $\eta^2_{\text{sub}} = 0.998$ a model is still $2{,}000\times$ above the threshold. Because the task is prefix products and the models are causal, accuracy at timestep ℓ of one long sequence equals accuracy for length ℓ, which we verified against literal per-length evaluation at horizons 2,000 and 20,000.

Table 5: No model passes, out of 57,920 measured (95% upper bound on the rate: 6.6×10^{-5}, the upper end of a Wilson interval for a pooled design, not a universal training rate). Entries are median $1 - \eta^2_{\text{sub}}$ per cell; the threshold is 10^{-6}. Two thirds of the 81 pooled arms include a soft orthogonality penalty, which helps one architecture and hurts the other, so the pooled rows should not be read as a comparison of architectures.

population	linear arm	LSTM arm	above threshold
pre-registered baseline arm	6.3–6.8×10^{-1}	9.7×10^{-2}–3.2×10^{-1}	10^5–$10^6\times$
all 81 arms, width 24	7.05×10^{-2}	5.56×10^{-1}	7.1–56×10^4
all 81 arms, width 48	4.54×10^{-2}	6.20×10^{-1}	4.5–62×10^4
best single measured model	$\mathbf{4.98 \times 10^{-5}}$ (linear, w48)		$50\times$

Table 6: Replication: Campaign B re-measurement on a corpus retrained from scratch (2,112 models). No unpenalised model passes; the Wilson 95% upper bound per 48-model cell is 0.074.

arm	width	n	converged	passes	median η^2_{sub}
linear baseline	24 / 48	192 / 192	0.927 / 0.943	0 / 0	0.367 / 0.327
LSTM baseline	24 / 48	192 / 192	0.995 / 1.000	0 / 0	0.874 / 0.851
orthogonal by construction (S_3, K_4)	24 / 48	96 / 96	1.000 / 1.000	0 / 0	0.9992 / 0.9996
+ relation penalty, $\beta \in \{10^{-3}, 0.1, 1\}$	24 / 48	576 / 576	1.000 / 1.000	**315 / 344**	1.0000 / 1.0000

Replication. Campaign B retrains the core cells from scratch (Table 6): none of 768 baseline models or 192 orthogonal ones passes, the typical values of η^2_{sub} match those of Campaign A, and under the penalty 88 + 98 D_4, 107 + 129 K_4, 32 + 21 S_3 and 88 + 96 $\mathbb{Z}_6$ models of 144 pass at widths 24 + 48. Over the 360 models of Campaign B's 10,000- and 20,000-step extensions,

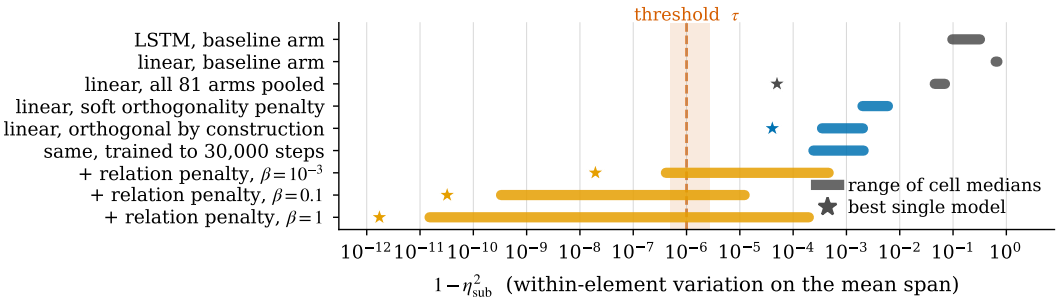


Figure 3: State consistency on the mean span, across training arms. $1 - \eta^2_{\text{sub}}$ is the within-element share of the state's variance on the span of the per-element means (§2). Rows are arms pooled over four groups; bars span the medians of an arm's cells, stars mark the best single model. τ was fixed on planted representations before any model was measured; only the relation-penalty rows cross it. A pass concerns the mean span, not the whole state; the defect on V is in Figure 7.

112 (31%) reached accuracy 0.99 and later fell below it, all of them in the un-penalised linear and LSTM arms and none in the orthogonal or penalised arms, which is why convergence is required at every logged step of the final quarter from Campaign B onwards. The readout sees the remainder: $\|W_{\text{out}} P_{\text{resid}}\|_F / \|W_{\text{out}}\|_F$ against 352 random projectors of the same rank gives median z by arm of 0.00 (linear baseline), $+4.17$ (LSTM baseline), -0.77 (hard-orthogonal) and $+1.58$ (relation penalty).

How much of the gap is scale? Splitting each state into a direction and a norm makes the scale caveat quantitative. On the whole state, the share of variance explained by the group element is 0.25 on raw states and 0.85 once each state is normalised to unit length for the linear baseline, 0.61 and 0.62 for the LSTM, and 0.973 both ways for the orthogonal arm, as it must be. Most of the linear baseline's distance from a representation is therefore magnitude. What remains after normalisation, about 0.15 of the variance for the baseline, 0.03 under orthogonality and less than 10^{-6} under the relation penalty, is what the paper is about; we could not identify what that path-dependent remainder encodes.

Ten times the budget. We trained the orthogonal arm, on which scale is controlled and accuracy is stable, for 30,000 steps at a constant learning rate on S_3 and K_4, with and without weight decay (Campaign C, 8 seeds per cell; Figure 4). No model passes at any of the 30 checkpoints; the best cell median anywhere along the trajectory is 2.4×10^{-4}, still $240\times$ above the threshold. The curves are not flat: every cell improves slowly until 15,000 to 20,000 steps and then gets worse, by a factor of two to five, ending roughly where it stood at step 3,000. Length generalisation does not improve either: L_{break} after 30,000 steps is between 118 and 410. The same rise after a minimum appears under the relation penalty in Appendix I, where a learning-rate schedule removes it; this control has not been rerun with a schedule. The 60,000- and 100,000-step runs of the unconstrained linear model and the LSTM are in Appendix E.

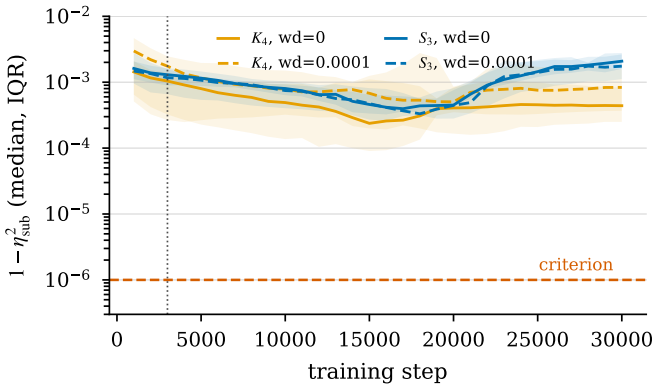


Figure 4: Orthogonal arm without relation penalty, trained for 30,000 steps (median and interquartile range of $1 - \eta^2_{\text{sub}}$ over 8 seeds; the dotted line is the original budget). No model crosses the threshold at any point, and every cell degrades after a minimum at 15,000–20,000 steps.

E CAMPAIGN E: STRUCTURED WORDS, COVERS, SPECTRA, AND THE UNCONSTRAINED ARM

Campaign E re-measures the stored checkpoints of Campaigns B–D and trains two new arms. Its two behavioural predictions (in paraphrase, P1: models trained on the defining relators of D_4, K_4, S_3, and on the redundant words of D_4, survive every family; P2: the redundant-word models of K_4 and $\mathbb{Z}_6$ break below 5,000 on $(ab)^n$ and a^n while censored under random words) were written and hashed before the structured-word test ran, and the rule combining them is met over all stored models and over converged models only; their registered text is given below, under "Registered predictions".

What was measured. On the stored checkpoints of Campaigns B–D: the five word families to 20,000 steps in float64 (§2); coset enumeration of the redundant words (sympy, cap 10^6 cosets); the mean-span screen with respect to the group presented by those words, with a null that relabels the states by 50 random refinements of the G partition with the same bucket sizes; the relator eigenvalues on V (a block on which a relator has eigenvalue -1, within angle 0.1, is a candidate signature of a central involution; the detector tests neither joint invariance nor centrality); the eigen-decomposition of the travelling element's operator and the spectral margin curve; patching to $t_0 + 1{,}000$; the Hessian of the penalty and of the joint loss at trapped S_3 checkpoints (Lanczos in the coordinates of the training parametrisation $W \mapsto \exp(W - W^\top)$, gauge directions projected out; fixture: the explicit $SO(6)$ construction gives gradient 4.6×10^{-14} and smallest non-gauge eigenvalue $+19.36$ in those coordinates, which are not the angle coordinates of Appendix K, where the trapped plane's own 2×2 Hessian has eigenvalues 16.378 and 45.590; the two agree on positive definiteness and the number should be read in its own coordinates). Trained fresh: the unconstrained-plus-penalty arm (384 models, §6) and the long-training arms of §3. Every new estimator was validated on planted fixtures before touching a trained model. Compute: 5.5 GPU-hours on one laptop GPU.

The redundant-word arm as run. With tokens $0 = r$ (or a) and $1 = s$ (or b, or a^{-1} for $\mathbb{Z}_6$), the code penalised the lexicographically first identities of G outside $\mathcal{R}$ at matched count: S_3: $(sr)^2$, s^4, r^3s^2; D_4: $(sr)^2$, s^4, r^4s^2; K_4: a^4, a^2b^2, ab^2a; $\mathbb{Z}_6$: a^3a^{-3}, $a^{-1}a$. The lists are seed-independent. For S_3, D_4 and K_4 they are the first identities outside $\mathcal{R}$ in shortlex order with token 0 before token 1; for $\mathbb{Z}_6$ shortlex order would give a^2a^{-2} in place of a^3a^{-3}, and both lists present $\mathbb{Z}$, so the prediction does not depend on which was used. Coset enumeration terminates for S_3 (order 6, kernel trivial) and D_4 (order 16, kernel $\mathbb{Z}_2$) and does not terminate for K_4 and $\mathbb{Z}_6$ within the cap; for those two the group is read off the presentation: $a^2b^2 = e$ and $ab^2a = e$ both say $b^2 = a^{-2}$, so $\widehat{G} = \langle a, b \mid a^4,\ b^2 = a^2 \rangle$, the amalgam $\mathbb{Z}_4 *_{\mathbb{Z}_2} \mathbb{Z}_4$, infinite and virtually cyclic; $a^{-1}a = e$ makes a^{-1} the inverse and $a^3a^{-3} = e$ then holds identically, so $\widehat{G} = \langle a \rangle \cong \mathbb{Z}$.

Table 1 shows that the penalty was met to 10^{-3} in only a minority of models outside K_4 (the words-met column uses the unweighted restricted defect on V of Appendix B.3). For D_4, r^4 and s^2 carry a -1 block in 48 of 48 models and $(rs)^2$ in 3, the SD_{16} pattern in 45; single relators with a -1 eigenvalue also occur in the un-penalised controls (14, 27 and 19 of 48), which is why the joint test of Appendix F is the evidence, not the counts. Labelling the states by the cover lowers the within-element residual by a factor 14–19 ($z = 29$ against random refinements) and 25% of models fall below 10^{-6} with respect to it; these residuals were computed on the reachable span, not with the mean-span screen, and with 8 words per element where Campaign E's registration specified 64 (Figure 5b). For K_4, a^2 and b^2 act as -1 on a block of median dimension 18.5 of 29 in every model and $(ab)^2$ does not, consistent with a, b of order 4 and ab of infinite order in $\widehat{G}$. The un-penalised orthogonal S_3 and the redundant-word $\mathbb{Z}_6$ models carry no signature, and every trapped S_3 model gives the trivial sign pattern.

Table 7 gives the per-family medians behind Figure 5a for every arm and group.

Random words and infinite covers. For the exact infinite cyclic presentation of $\mathbb{Z}_6$'s redundant words, uniformly sampled a and a^{-1} induce a centred random walk: the signed exponent after T tokens has variance T, whereas a^T has exponent T. If failure occurs when a fixed accumulated phase is reached, this suggests horizons scaling as δ^{-2} under random words and δ^{-1} under the travelling word; a related argument applies to a centred diffusive coordinate in the virtually cyclic K_4 presentation. Infinite order alone does not imply centred diffusion, and approximate learned operators

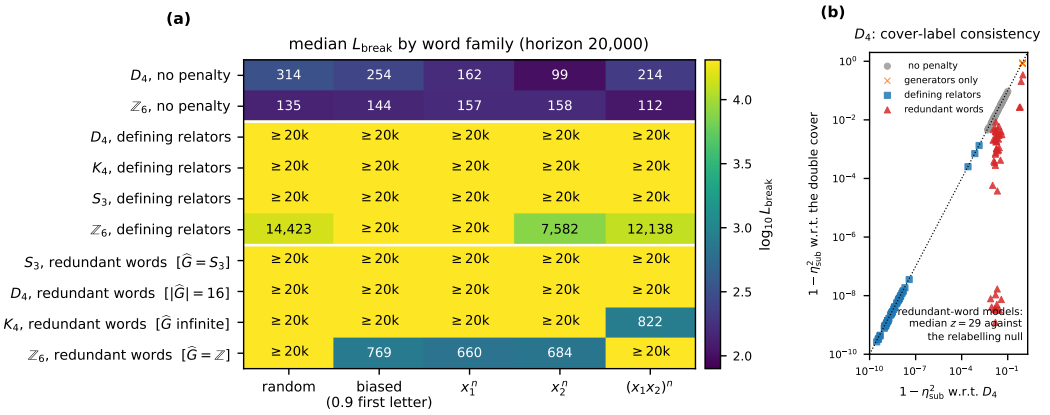


Figure 5: Redundant-word penalties under structured words (Campaign E). **(a)** Median L_{break} by word family on Campaign D's orthogonal checkpoints, 48 models per row (both widths and L_{train}), float64 to 20,000 (Table 7 gives the censoring shares). **(b)** D_4, models trained at $L_{\text{train}} = 24$: the within-element residual on the reachable span (not the mean-span screen; 8 words per element) with respect to D_4, against the same residual with respect to the double cover.

need not reduce to that exact walk, so the comparison $822^2 \approx 7 \times 10^5 \gg 20{,}000$ is a heuristic, not a measured survival law. In the K_4 redundant-word row 73% of models are censored under random words (Table 7) and 6 of the 48 never converged, so some converged models break under random words before 20,000; we have not examined those models individually. The data establish a large gap between random and structured horizons, not its exponent.

Table 7: Median L_{break} by word family on Campaign D's orthogonal checkpoints, as re-measured in Campaign E (censoring share in parentheses; n models per row, both widths and both L_{train}). Deterministic families: first position beyond L_{train} at which accuracy over the preceding 100 positions falls below 0.90, so a value of 99, the first position at which the window is evaluated, means that the family fails within the first hundred steps (never-converged models fail immediately and enter the medians as early breaks: the evaluator applies no convergence gate, and a model already below 0.90 at L_{train} is recorded at $L_{\text{train}} + 1$, stored as 13 because Campaign E applied $L_{\text{train}} = 12$ to every model; the redundant-word rows contain 17 never-converged models, 4 D_4, 6 K_4 and 7 S_3, all at width 24); stochastic families (random, biased): 32 sequences. Campaign D's own evaluation of the same checkpoints (Table 18) gives 319 and 146 for the two control rows and 14,412 for the $\mathbb{Z}_6$ defining-relator row under random words, against 314, 135 and 14,423 here: two evaluation runs, not two populations.

arm	G	n	random	biased	x_1^n	x_2^n	$(x_1x_2)^n$
no penalty	D_4	48	314 (0%)	254 (0%)	162 (0%)	99 (0%)	214 (0%)
no penalty	$\mathbb{Z}_6$	48	135 (0%)	144 (0%)	157 (0%)	158 (0%)	112 (0%)
defining relators	D_4	48	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (98%)	≥20,000 (100%)
defining relators	K_4	48	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)
defining relators	S_3	48	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)
defining relators	$\mathbb{Z}_6$	48	14,423 (25%)	≥20,000 (88%)	≥20,000 (90%)	7,582 (8%)	12,138 (17%)
redundant words	D_4	48	≥20,000 (92%)	≥20,000 (92%)	≥20,000 (92%)	≥20,000 (92%)	≥20,000 (92%)
redundant words	K_4	48	≥20,000 (73%)	≥20,000 (85%)	≥20,000 (88%)	≥20,000 (88%)	822 (8%)
redundant words	S_3	48	≥20,000 (85%)	≥20,000 (85%)	≥20,000 (83%)	≥20,000 (85%)	≥20,000 (85%)
redundant words	$\mathbb{Z}_6$	48	≥20,000 (92%)	769 (0%)	660 (0%)	684 (0%)	≥20,000 (100%)

Spectra and the replay. For every converged redundant-word model of K_4 (family $(ab)^n$, $M = A_bA_a$) and $\mathbb{Z}_6$ (family a^n, $M = A_a$), the eigen-decomposition of M on the full space gives per eigenplane the angle error δ_j from the nearest ideal angle ($\{0, \pi\}$ and multiples of $\pi/3$) and the logit amplitude $a_j = \|W_{\text{out}}^\top \text{Re}(c_j v_j)\|$ where $h_0 = \sum_j c_j v_j$. The margin at block endpoints is computed from $\ell(n) = W_{\text{out}} \sum_j R_j(n\varphi_j)h_j + b$, and inside a block of length q from $\ell(nq + r) = W_{\text{out}} A_{x_1 \cdots x_r} M^n h_0 + b$; the token-wise correctness sequence this implies reproduces the simulated L_{break} of all 90 models exactly under L_{break}'s own rolling-window definition (the first negative endpoint margin alone lands within the 100-position window in 82 of 90). This is the forward pass written in the eigenbasis of M, so the agreement is a check of the decomposition and of the eigensolver, not a test of a mechanism; the mechanistic content is the split of the planes into drifting and bounded ones below. Reference spectra: models trained on the defining relators have amplitude-weighted angle errors of 2×10^{-5} (K_4) and 4×10^{-6} ($\mathbb{Z}_6$); the un-penalised orthogonal controls 1.9×10^{-2}; the redundant-word models 4×10^{-2} (K_4) and 1.2×10^{-2} ($\mathbb{Z}_6$) overall, but 10^{-3} in the

two dominant planes, which carry about half the logit amplitude each, with a low-amplitude tail at 0.5. A one-parameter law, break length $\propto 1/\delta$, holds only loosely (Spearman ≥ 0.7 for $\mathbb{Z}_6$ with the dominant-plane error, not for K_4). Table 8 summarises the spectra per arm and group.

Table 8: Spectral summary per arm and group: number of models, amplitude-weighted angle error δ_{eff}, error of the dominant plane δ_{top}, and median angle error over all planes. The simulated break medians of the two redundant-word rows are in Table 7 and §4.

arm	G	n	δ_{eff}	δ_{top}	median δ_j
defining relators	K_4	48	2.0e-05	2.5e-05	0.00
defining relators	$\mathbb{Z}_6$	48	3.9e-06	5.1e-06	0.00
no penalty	$\mathbb{Z}_6$	48	1.9e-02	6.5e-03	0.21
redundant words	K_4	42	4.0e-02	9.5e-04	0.41
redundant words	$\mathbb{Z}_6$	48	1.2e-02	1.3e-03	0.16

The trapped plane, measured on the checkpoints. On eight trapped S_3 checkpoints of Campaign C ($\beta = 0.1$, width 24), restricting the three relators to the orbit span gives a single 2-plane on which r^3 acts as a rotation by 0.534 and s^2, $(rs)^2$ by 0.869, the values $3\alpha^\star$ and $2\theta^\star$ of the closed form (Appendix K), and the identity elsewhere. That plane holds 2–7% of the state's energy and 14–28% of the readout's weight, so it is a candidate for $\mathcal{O}$ in Proposition 2; placing it there at every length needs $B > 0$ from absolute, phase-wise margins (Remark 2). The joint-loss Hessian at three such checkpoints has most negative non-gauge eigenvalue -2.1×10^{-5}, -2.0×10^{-5} and -1.6×10^{-4} at gradient norms 3×10^{-4} to 1.8×10^{-3}: a slow region, not a critical point, with reciprocal-curvature scales $1/(\eta|\lambda_{\min}|)$ of 1.6×10^7, 1.7×10^7 and 2.1×10^6 steps at $\eta = 3 \times 10^{-3}$ (plain-gradient scales, not escape-time estimates for Adam). The penalty alone at the same checkpoints has $\lambda_{\min}$ within $\pm 5 \times 10^{-5}$ of zero, with the next eigenvalues at 11–19: the fixture's $+19.36$ is the smallest non-gauge curvature at the trap in the coordinates of the training parametrisation, and the trained models sit on its flat approach.

The unconstrained arm. Tables 9 and 10 give the unconstrained-plus-penalty arm in full. Under a cosine decay to zero within the same 3,000 steps, at $\beta = 0.1$, the unconstrained arm passes no model in any group; that schedule halves the integrated learning rate, and matching the integral restores most of the constant-rate result (Appendix I). And 18 of the 48 S_3 models at $\beta \geq 0.1$ have $\dim V = 7$ rather than 6 and sit in a real local minimum of the penalty. One direction of V is an eigenvector of A_r with eigenvalue $a^\star = -0.496151$ and nearly one of A_s with eigenvalue $b^\star = 1.08396$; $(a^\star, b^\star)$ minimises $(a^3 - 1)^2 + (b^2 - 1)^2 + ((ab)^2 - 1)^2$ locally (Hessian eigenvalues 6.26 and 10.86), predicts the three per-relator defects to five digits, and gives A_{r^3} the eigenvalue -0.122 on that direction. Campaign E had read this eigenvalue as a -1 signature of a central extension, because its counter tested the angle only; there is none, and we retract that reading. These 18 models break within a hundred steps on s^n, along which the direction is scaled by about 1.084 per step, while surviving 20,000 random steps; 15 of them pass the screen, the direction lying outside the span of the per-element means, although by Lemma 1 their state is not a function of the S_3 element. Nine models under the cosine schedule sit in the same minimum; orthogonal operators cannot, since it needs a real eigenvalue above 1.

Long training. The unconstrained linear arm ran to 60,000 steps (6 seeds, four groups, weight decay 0 and 10^{-2}) and the LSTM to 100,000 (8 seeds, D_4, S_3, $\mathbb{Z}_6$, three weight-decay levels), cosine decay to zero, twenty log-spaced measurement points with weights saved at each (Figure 6, Table 11). The registered rule returns "implicit bias does not supply exactness" in 13 of 17 cells and "unresolved" in the four linear cells with weight decay, whose slope (-1.0 to -1.3 in $\log(1 - \eta_{\text{sub}}^2)$ against $\log \text{step}$) exceeds the -0.5 the rule uses; these registered labels summarise finite-budget behaviour, and a log–log extrapolation far beyond the measured range is not evidence about eventual convergence. Two limits of the design: the orthogonal arm was not run to 100,000 steps, so the schedule-plus-long-run control for that arm is missing; and the training code applies weight decay with the configured learning rate rather than the decayed one, so under cosine decay the middle level (10^{-4}) shrinks the weights by 3% over the whole run and is inert, which makes the weight-decay contrast effectively two-level.

The LSTM and the linear baseline under structured words. Without any penalty the LSTM shows a gap between random and structured words: at $L_{\text{train}} = 12$ its median s^n break on D_4 and

Table 9: Unconstrained linear RNN with the relation penalty (Campaign E; width 24, $L_{\text{train}} = 24$, 24 seeds per cell, 3,000 steps; the cosine rows decay the learning rate to zero within those 3,000 steps): mean-span passes and defects. "pass (orth.)" is the matched orthogonal-plus-penalty cell at the constant rate (Campaigns A and C); no orthogonal cell is matched to the cosine rows, whose orthogonal comparison in Appendix I decays over 10,000 steps. Defect: visit-weighted, median. $\|A\|_F$: median Frobenius norm of the operators (4.90 for an orthogonal 24×24 matrix).

β, schedule	G	pass	pass (orth.)	defect	$\|A\|_F$
10^{-3}, const	D_4	0.00	0.00	2.6e-01	5.34
10^{-3}, const	K_4	0.00	0.21	3.7e-01	4.77
10^{-3}, const	S_3	0.00	0.00	3.8e-01	4.89
10^{-3}, const	$\mathbb{Z}_6$	0.00	0.00	4.7e-01	4.78
0.1, const	D_4	1.00	0.88	2.7e-04	5.57
0.1, const	K_4	0.54	1.00	5.3e-04	5.41
0.1, const	S_3	0.96	0.50	3.8e-04	5.53
0.1, const	$\mathbb{Z}_6$	0.00	0.67	1.6e-03	5.52
1, const	D_4	1.00	0.92	4.3e-05	5.40
1, const	K_4	1.00	1.00	7.1e-05	5.38
1, const	S_3	0.92	0.25	5.7e-05	5.44
1, const	$\mathbb{Z}_6$	0.58	1.00	2.3e-04	5.44
0.1, cosine	D_4	0.00	—	9.9e-04	5.55
0.1, cosine	K_4	0.00	—	1.9e-03	5.41
0.1, cosine	S_3	0.00	—	1.6e-03	5.51
0.1, cosine	$\mathbb{Z}_6$	0.00	—	6.1e-03	5.50

Table 10: The same arm as Table 9: median L_{break} by word family (censoring share in parentheses), evaluated in float64 to 20,000. A value of 99 is the floor of the windowed definition (Table 7).

β, schedule	G	random	biased	x_1^n	x_2^n	$(x_1x_2)^n$
10^{-3}, const	D_4	196 (0%)	442 (0%)	406 (0%)	318 (0%)	417 (21%)
10^{-3}, const	K_4	196 (0%)	113 (0%)	328 (8%)	182 (8%)	239 (0%)
10^{-3}, const	S_3	224 (0%)	170 (0%)	178 (4%)	138 (4%)	477 (4%)
10^{-3}, const	$\mathbb{Z}_6$	62 (0%)	164 (0%)	344 (0%)	99 (0%)	4,629 (0%)
0.1, const	D_4	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)
0.1, const	K_4	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)
0.1, const	S_3	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (58%)	≥20,000 (100%)
0.1, const	$\mathbb{Z}_6$	4,807 (0%)	≥20,000 (62%)	≥20,000 (100%)	2,530 (0%)	≥20,000 (79%)
1, const	D_4	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)
1, const	K_4	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)
1, const	S_3	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (62%)	≥20,000 (100%)
1, const	$\mathbb{Z}_6$	≥20,000 (88%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (67%)	≥20,000 (100%)
0.1, cosine	D_4	≥20,000 (58%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (96%)	≥20,000 (96%)
0.1, cosine	K_4	≥20,000 (88%)	≥20,000 (88%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (92%)
0.1, cosine	S_3	≥20,000 (96%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (58%)	≥20,000 (100%)
0.1, cosine	$\mathbb{Z}_6$	1,699 (0%)	8,740 (4%)	≥20,000 (100%)	1,025 (0%)	≥20,000 (88%)

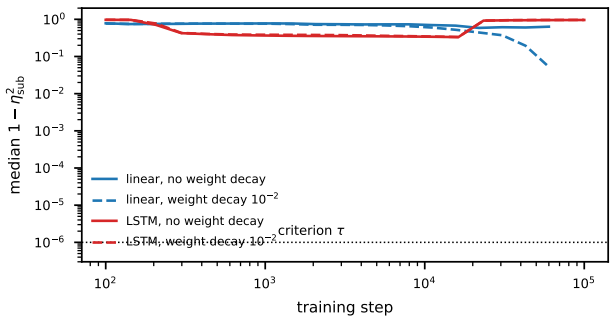


Figure 6: Long training under a cosine schedule: median $1 - \eta^2_{\text{sub}}$ against step. With weight decay the unconstrained linear residual falls roughly as a power law over the measured range and is still 4.7 orders of magnitude above τ at 60,000 steps; the LSTM's residual grows while its accuracy stays at 1.0000.

S_3 is within a hundred steps (40% and 42% of the models survive s^n to 20,000) against random-word medians in the thousands, and 44% and 54% of those models (23% and 31% at $L_{\text{train}} = 24$) break

Table 11: Long training: median $1 - \eta^2_{\mathrm{sub}}$ at 3,000 steps and at the final step, and median held-out accuracy at the final step.

arch.	weight decay	models	final step	$1 - \eta^2_{\mathrm{sub}}$ at 3k	at final	held-out acc.
linear	0	24	60,000	0.737	0.633	1.0000
linear	0.01	24	60,000	0.711	0.053	1.0000
LSTM	0	24	100,000	0.352	0.959	1.0000
LSTM	0.0001	24	100,000	0.373	0.961	1.0000
LSTM	0.01	24	100,000	0.376	0.964	1.0000

on their worst structured family more than ten times earlier than on random words. The gap depends on the training run: Campaign F retrained 48 such LSTMs with the same optimiser, rate, batch, pool size and schedule but different seeds and a pool drawn at length 12, and 18 of 48 broke on s^n, against 34 of 48 here (Appendix F). Table 12 gives the per-cell results. The registered rule for the LSTM (its worst structured family stays within a factor of two of its random-word median) is met in every group, but it pools the two training lengths, and at $L_{\mathrm{train}} = 24$ nearly every LSTM is censored under every family, which carries the pooled verdict: the rule is met by censoring and is not cited as evidence in the main text. The split by training length is where the failures are.

Table 12: Block-0 LSTM and unconstrained linear baselines under the five families, per group and training length: median L_{break} (censoring share), and the share of models whose worst structured family breaks more than $10\times$ earlier than random words. A value of 99 is the floor of the windowed definition (Table 7): the family fails within the first hundred steps.

arch.	G	L_{train}	n	random	biased	x_1^n	x_2^n	$(x_1x_2)^n$	worst $< 0.1\times$random
LSTM	D_4	12	48	5,683 (33%)	≥20,000 (94%)	≥20,000 (98%)	99 (40%)	≥20,000 (96%)	44%
LSTM	D_4	24	48	≥20,000 (92%)	≥20,000 (100%)	≥20,000 (100%)	≥20,000 (77%)	≥20,000 (100%)	23%
LSTM	K_4	12	48	≥20,000 (73%)	≥20,000 (65%)	≥20,000 (94%)	≥20,000 (94%)	≥20,000 (100%)	33%
LSTM	K_4	24	48	≥20,000 (98%)	≥20,000 (67%)	≥20,000 (90%)	≥20,000 (90%)	≥20,000 (100%)	46%
LSTM	S_3	12	48	4,819 (27%)	≥20,000 (90%)	≥20,000 (94%)	99 (42%)	≥20,000 (100%)	54%
LSTM	S_3	24	48	≥20,000 (98%)	≥20,000 (94%)	≥20,000 (96%)	≥20,000 (73%)	≥20,000 (100%)	31%
LSTM	$\mathbb{Z}_6$	12	48	788 (23%)	≥20,000 (58%)	≥20,000 (92%)	≥20,000 (92%)	≥20,000 (69%)	19%
LSTM	$\mathbb{Z}_6$	24	47	≥20,000 (51%)	≥20,000 (55%)	≥20,000 (70%)	≥20,000 (70%)	≥20,000 (79%)	30%
linear	D_4	12	48	33 (0%)	25 (0%)	99 (0%)	99 (0%)	99 (0%)	0%
linear	D_4	24	46	62 (0%)	60 (0%)	99 (0%)	165 (0%)	103 (0%)	0%
linear	K_4	12	47	35 (0%)	39 (0%)	127 (0%)	144 (0%)	99 (0%)	0%
linear	K_4	24	41	66 (0%)	58 (0%)	309 (0%)	279 (0%)	99 (0%)	0%
linear	S_3	12	48	45 (0%)	44 (0%)	99 (0%)	99 (0%)	99 (0%)	0%
linear	S_3	24	39	88 (0%)	84 (0%)	99 (0%)	120 (0%)	146 (0%)	0%
linear	$\mathbb{Z}_6$	12	48	28 (0%)	30 (0%)	99 (0%)	99 (0%)	99 (0%)	0%
linear	$\mathbb{Z}_6$	24	42	54 (0%)	44 (0%)	99 (0%)	99 (0%)	99 (0%)	0%

Patching to a longer horizon. At $t_0 = L_{\mathrm{train}}/2$ the state is replaced by the state of a different word spelling the same element (donor length $\leq L_{\mathrm{train}}$), and the identical suffix is continued. Table 13 gives the LSTM's unpatched accuracy at L_{train}, $t_0 + 100$ and $t_0 + 1{,}000$ and the paired difference; eight seeds per cell, 256 sequences. The residual is behaviourally inert wherever the model is still accurate: the penalised and redundant-word orthogonal arms read 1.000 with a paired difference of $+0.0000$ at every horizon, and the un-penalised orthogonal arm, whose paired differences never exceed 0.021 in magnitude, is itself at chance by $t_0 + 1{,}000$, which is drift, not patching (the orthogonal rows are summarised here and not printed; the table gives the LSTM rows).

Table 13: Patching on the LSTM baseline: unpatched accuracy / paired difference (patched minus unpatched) at three horizons.

G	at L_{train}	at $t_0 + 100$	at $t_0 + 1{,}000$
D_4	1.000 / +0.0000	0.999 / +0.0000	0.996 / +0.0000
K_4	1.000 / +0.0000	1.000 / +0.0000	0.996 / +0.0000
S_3	1.000 / +0.0000	1.000 / +0.0000	0.999 / +0.0000
$\mathbb{Z}_6$	1.000 / +0.0000	0.977 / -0.0015	0.843 / +0.0010

Registered predictions. Campaign E's rule (`T9_amend2.md`, registered 2026-09-21T13:55:19Z, SHA-256 `730decf0bf1dc662322209125bafbab815a4853a797f40561ec713192059d363`), verbatim:

> **P1** (finite cover or exact: immune). Variant (i) on D4, K4, S3 and variant (iv) on D4: in at least 90% of converged models no family breaks before 20,000. **P2** (infinite cover: vulnerable). Variant (iv) on K4 under `alt`, and variant (iv) on Z6 under `power_a`: median

> L_break below 5,000, while the same models' median under `random` is censored at 20,000. **H-drift supported if P1 and P2 both hold; refuted if P2 fails for both K4 and Z6 (at least 90% of those models survive the structured family). Anything else: undetermined.** Report L_break per family, variant, group; no pooling across groups.

The registration fixes the models (both widths, all stored models), float64 evaluation to 20,000, the rolling rule for deterministic words and the five families. P1 and the refutation clause are per-model proportions; P2 uses medians. Evaluated as registered, H-drift is supported. Five facts qualify this. (1) The verdict was first reported as undetermined, P1 failing on the redundant-word D_4 row at 0.833, because the evaluation grouped models by index alone and merged the two widths; after that coding fix it is 0.917 over 48 models, and H-drift is supported. (2) P1's text says converged models, but the registered code evaluates all stored models. (3) A second rule restricted to converged models (`T9_rule2.md`, 2026-09-22T02:08:31Z, SHA-256 `11d821477ac7f93150acf53603f954cf45e06f22a3cece4147acd1fa9a748cd7`) was written after the first outcome was known; it is not prospective, and we report the first rule as the registered result. (4) Its convergence flag is the final-step accuracy, not every logged step of the final quarter as its text says. (5) The block0 relation-penalty models the rule names were never evaluated. Campaign F's registration avoids these five problems, though it left two choices unspecified (Appendix F).

F Campaign F: registered tests

Registration. The predictions were registered on 23 September 2026 at 07:06:45 UTC in `F_predictions.md` (SHA-256 `10529bf2e1f54163834c524f5b504f9c9ba9446468407e1a4aa767eecc57be3a`), 89 seconds before the first evaluation started and before Campaign F's evaluation scripts were written (they reuse Campaign E's evaluator functions). The timing rests on our own git history; no third-party timestamp was taken. The file states each prediction and decision rule verbatim; training of the 720 new models was allowed to precede it, evaluation was not. Every group-order claim the predictions use was checked beforehand in exact arithmetic (Knuth–Bendix completion, Todd–Coxeter enumeration, explicit homomorphisms to $\mathbb{Z}$, D_∞ and finite groups): 81 of 82 checks pass, and the exception is not an order claim but the observation that F1b's three parity controls coincide with the D_4 labelling, so F1b compares four distinct labellings, not six. Evaluation is in float64, every model uses its own L_{train}, and every seed is recorded. Table 14 lists the 19 registered predictions: 17 pass, one fails, and one (F1d-1) is indeterminate, since its comparator's span was not registered and the other reading fails. They carry different weight. Most are consistency checks that follow from the measured defects once the imposed identities hold (Corollary 1) or from Proposition 2: F1a-1 to F1a-3, F1c-1 to F1c-3 and F2a-1 to F2a-3; their value lies in the point estimates (a ratio of 1.958 against 2, certificates within a factor of 2.9). The outcomes the setup does not force are that training realises a cover rather than the target group, which satisfies the same identities: the D_4 sign pattern (F1b-1; D_4's own representation would have avoided it, although Campaign E had partly seen it on the same checkpoints) and the moved breaks on newly trained models (F2b). F1b-2 and F1b-3 check instruments.

Table 14: Campaign F: every registered prediction and its outcome.

test	prediction	outcome	verdict
F1a-1	K_4: $(aba)^n$, $(bab)^n$, $(a^2b)^n$ censored at 20,000	42/42 models, all three	pass
F1a-2	K_4: $(ab^3)^n$ breaks, per-model ratio to $(ab)^n$ in $[1.5, 3]$	median 1.958 (IQR 1.943–1.980)	pass
F1a-3	$\mathbb{Z}_6$: $(a^k a^{-k})^n$ censored; $k \neq j$ at $(k+j)/\lvert k-j\rvert$, ratio in $[0.5, 2]$	14/14 censored; ratios 0.965–0.994	pass
F1b-1	D_4: $E_{r^4} = E_{s^2}$, no -1 of $(rs)^2$ there, block invariant	angle 3.8×10^{-5}; 0 in 48/48; residual 8.0×10^{-5}	pass
F1b-2	SD_{16} labelling lowest in $\geq 75\%$ of models	43/48 (mean span); 48/48 (reachable span)	pass
F1b-3	detector false positives $< 5\%$	0/1,000 random pairs; 0/48 controls	pass
F1c-1	snapping the dominant planes: censored in $\geq 75\%$	90/90, also to 10^6	pass
F1c-2	snapping the tail: break within $\pm 25\%$	median ratio 1.003	pass
F1c-3	snapping two random planes: median within $\pm 25\%$	0.996	pass
F1d-1	$\lvert\rho(d_{\text{obs}})\rvert \geq \lvert\rho(\text{unweighted defect})\rvert$	$0.516 \geq 0.486$; $0.516 < 0.519$ on the 64-word span	indeterminate
F2a-1	x^n, y^n, $(xyx)^n$, $(yxy)^n$ censored in $\geq 95\%$	197/197 (by construction)	pass
F2a-2	$\geq 90\%$ of breaking models break first on $(xy)^n$	183/183	pass
F2a-3	Spearman(L_{break} on $(xy)^n$, angle error) ≤ -0.7, pooled	-0.971 (-0.999 within groups)	pass
F2b	$K_4\{a^2, b^2\}$: $(ab)^n$ breaks below random/10	2,484 against 2,000	**fail**
F2b	$K_4\{a^2, (ab)^2\}$: b^n breaks, a^n, $(ab)^n$, $(bab)^n$ survive	191; censored	pass
F2b	$K_4\{b^2, (ab)^2\}$: a^n breaks, b^n, $(ab)^n$, $(aba)^n$ survive	216; censored	pass
F2b	$K_4\{a^2, b^2, (ab)^4\}$: all survive	censored	pass
F2b	$D_4\{s^2, (rs)^2\}$: r^n breaks, s^n, $(rs)^n$, $(rsr)^n$, $(r^2s)^n$ survive	344; censored	pass
F2b	$D_4\{r^8, s^2, (rs)^2\}$: all survive	censored	pass

F1a: discriminating words. Stored redundant-word checkpoints, Campaign D. Of the 48 K_4 models, 42 meet their penalised words (defect below 10^{-3} on V), exactly the converged ones. In all 42, $(aba)^n$, $(bab)^n$ and $(a^2b)^n$, of order 4 in $\widehat{G}$, survive 20,000 tokens, while $(ab)^n$ breaks at a median of 954.5 and $(ab^3)^n$ at 1,885.5. The ratio's IQR excludes 2 by about 2%; we did not model how the 100-token window's offset differs between word lengths, and the ratio of first errors was not registered. The decision rule, which would have rejected "the model represents $\widehat{G}$" had an order-4 word broken within ten times the $(ab)^n$ break (9,545), is not triggered. On $\mathbb{Z}_6$ only 14 of 48 converged models meet their words (below); in them a^n breaks at a median of 619.5 and the four $k \neq j$ families at observed/predicted ratios of 0.972, 0.979, 0.965 and 0.994 for $(k, j) = (2, 1), (3, 1), (3, 2), (1, 2)$.

A trap in the redundant-word penalty. The other 34 $\mathbb{Z}_6$ models sit in a local minimum of the penalty: the operator of $a^{-1}a$ rotates one plane (two in two models) by $\theta^\star = 114.0948°$, every other angle being below $0.0073°$. On a plane where A_a and $A_{a^{-1}}$ commute the penalty is $2[(2 - 2\cos\theta) + (2 - 2\cos 3\theta)]$, whose non-zero critical point $\sin\theta + 3\sin 3\theta = 0$ is a minimum (second derivative 32.66) and predicts the measured parameter defects to six digits. These models still keep every $(a^k a^{-k})^n$ unbroken to 20,000.

F1b: the D_4 cover. On the 48 D_4 redundant-word models the -1 eigenspaces of r^4 and s^2 coincide (median largest principal angle 3.8×10^{-5} rad), are invariant under A_r and A_s (median residual 8.0×10^{-5}), and carry no -1 eigenvalue of $(rs)^2$ in any model; all three hold in 44 of 48, the 44 converged. The detector fires in none of 1,000 Haar-random pairs in $O(24)$ or $SO(24)$, none of the 48 un-penalised controls and none of the 48 defining-relator models; random pairs and controls often have -1 eigenvalues of r^4 and s^2 separately, never on a common block. Labelled by each order-16 cover, the states fit SD_{16} best in 43 of 48 models under the registered mean-span screen and in 48 of 48 under the reachable-span residual; 12 models pass the 10^{-6} screen under SD_{16} and none under D_{16}, Q_{16} or D_4. The mean span cannot tell a cover from its quotient when the sampled states are balanced over the two lifts (on a planted exact SD_{16} fixture every labelling's residual is below 6×10^{-30}); it separates them here only because its random draw is not balanced, which is why the reachable-span residual is the stronger evidence.

F1c: plane surgery and Proposition 2. The 90 redundant-word models of Remark 1 (42 K_4, 48 $\mathbb{Z}_6$, including the 34 trapped $\mathbb{Z}_6$ models; K_4 alone meets every rule). The operator of the block is split into its rotation planes by a real Schur form (block diagonal to 7.8×10^{-14}), and a plane is snapped by replacing its rotation with the ideal one; for K_4 the edited product is realised as $A'_b = M' A_a^\top$ with A_a kept. The two dominant planes carry a median 0.92 of the readout-visible amplitude, at angle errors of 1.15×10^{-3} and 0.99×10^{-3}. Snapping them leaves 90 of 90 models unbroken to 20,000 tokens and to 10^6; snapping every other plane changes the break by a median per-model factor of 1.003 (IQR 0.999–1.015), and two random tail planes by 0.996 over 450 draws. Table 15 gives B and C of Proposition 2 with the dominant planes as $\mathcal{D}$ and prefix phases included in μ^0 and g_j. The bound $\mu \geq B - nC$ holds at every one of 10^6 positions in all 720 (model, condition) rows, as it must; the finding is its tightness. The edited operators were evaluated on the targeted word only: the K_4 edit changes A_b, and its effect on other words and on training-length accuracy was not measured.

Table 15: F1c: certified and observed horizons, in blocks of the periodic word.

	n	$B > 0$	median μ^0	median B	median B/C	median observed	observed/(B/C) [min]
K_4	42	40	9.76	5.70	114.5	477.8	2.93 [2.03]
$\mathbb{Z}_6$	48	47	9.90	6.72	266.4	661.0	2.77 [1.63]
all	90	87	9.84	6.21	184.9	595.5	2.87 [1.63]

F1d: the exact span and the observable part. The Krylov closure of h_0 under the operators, at relative tolerance 10^{-10}, is the whole state space in all 2,088 stored models, so the exact orbit span is $\mathbb{R}^H$ and the V of this paper is the span the states occupy, directions visited at 10^{-5} to 10^{-7} of the largest singular value excepted. The minimum restricted defect over the 756 ordinarily trained models is 0.972441 on both spans, the same model. The observable part, the Krylov closure of the readout's rows under the transposed operators, is also the whole space in 271 of 271 models, so the defect on it is the parameter defect over $\sqrt{H}$. Its Spearman correlation with L_{break} over the 271 uncensored Campaign C models is -0.516 [-0.616, -0.405], above the registered comparator (-0.486, the

unweighted defect on the 8-word sampled V) but below the same defect on the 64-word V (-0.519); the pass is fragile, and all trail $1 - \eta^2_{\text{sub}}$ (-0.803) and the visit-weighted defect (-0.859 [-0.894, -0.813]).

F1e: horizons beyond 20,000. For periodic words the break follows from the eigendecomposition of the block operator, with every prefix phase, to 10^7 tokens. It agrees with direct float64 rollout wherever the break falls below 20,000 (1,009 of 1,009 rows whose rollout stays finite; the 57 disagreements are float64 overflows of unconstrained models) and with an eigen-free blocked rollout beyond it (2,639 of 2,639 rows). All 1,770 orthogonal periodic rows censored at 20,000 break before 10^7 tokens, the longest at 8,003,069. Table 16 gives the defining-relator models; the redundant-word models' censored periodic words extend to medians of 37,986 to 199,015, except the travelling words, 954 (K_4) and 660 and 684 ($\mathbb{Z}_6$). Random words are not extended.

Table 16: F1e: defining-relator models (Table 7's populations), median break on periodic words extended to 10^7 tokens, against the un-penalised orthogonal controls.

G	word	median break	control median	ratio
D_4	r^n, $(rs)^n$, s^n	196,944; 118,360; 55,896	162; 214; 99	1,220; 554; 565
K_4	a^n, $(ab)^n$, b^n	92,111; 126,214; 74,444	99; 292; 99	930; 432; 752
S_3	r^n, $(rs)^n$, s^n	144,096; 179,129; 74,496	172; 224; 99	840; 798; 753
$\mathbb{Z}_6$	a^n, $(aa^{-1})^n$, $(a^{-1})^n$	266,676; 12,138; 7,582	157; 112; 158	1,699; 108; 48

F2a: single reflections. 288 models (three groups, widths 24 and 48, L_{train} 12 and 24, 24 seeds), no penalty. K_4 and S_3 converge in 96 of 96; D_4 in 5 of 96, stuck from step 500 at a common accuracy plateau (0.955 at width 24 and 0.962 at width 48 for $L_{\text{train}} = 12$) that a 10,000-step run does not leave, although D_4 generated by two reflections is exactly representable; we have not established why. $A_x^2 = I$ holds to rounding, so the involution words cannot break and P-F2a-1 confirms the construction, not a learned property. Of the 197 converged models, 183 break before 20,000, each first on $(xy)^n$; the median $(xy)^n$ break is 4,542 (K_4 4,309, S_3 4,850, D_4 498, $n = 5$), random words are censored under both estimators, and the Spearman correlation with the angle error is -0.999 within K_4 and within S_3. The accumulated phase error at the break has median 0.777, 0.520 and 0.397 rad on K_4, S_3 and D_4 ($n = 5$), against $\pi/(2m) = 0.785$, 0.524 and 0.393, half the spacing of the dihedral orbit (unregistered).

F2b: subsets of the relators. 144 models, orthogonal arm, $\beta = 0.1$, 24 seeds per set (Figure 1). 27 of 28 registered family predictions hold. In the failed set, $\{a^2, b^2\}$ on K_4, $(ab)^n$ breaks in 22 of 24 models at a median of 2,484, the only family that breaks, but not below one tenth of the censored random-word median. Two families we should have registered as breaks do break: under $\{a^2, (ab)^2\}$, $aba = b^{-1}$ travels (median 419), and under $\{b^2, (ab)^2\}$, $bab = a^{-1}$ does (496); their breaks are 2.2 and 2.3 times those of b^n and a^n, below the speed ratio of 3, which we have not explained. In the two finite-cover sets the penalty is met in only 14 and 5 of 24 models, the others resting at recurring stationary points of the penalty, yet every family survives, as it does for the S_3 redundant words (met in no model): the unmet words leave bounded components, as in Remark 2, not drifting ones.

F2c: LSTM coverage. 96 LSTMs on S_3 and D_4 at $L_{\text{train}} = 12$, half trained with 20% of the pool replaced by runs of a single generator. The control did not reproduce Campaign E's failure: 18 of 48 break on s^n, against 34 of 48 in Campaign E's LSTMs, and its median is censored, so the registered condition for "coverage" (augmented median censored or ten times the control's) holds only through its censored branch and says nothing. The runs differ in seeds and in the length at which the pool was drawn; the cause of the difference is not established.

F2d: schedule controls. 192 models at $\beta = 0.1$, four groups, 24 seeds. The 3,000-step cosine of Campaign E halves the integrated learning rate. With the integral matched (peak 6×10^{-3}), the unconstrained arm passes the screen in 40 of 96 models, has a defect below 10^{-3} on V in 65 and survives 20,000 random steps in 74, against 0, 23 and 58 under Campaign E's cosine and 60, 70 and 72 at the constant rate. The orthogonal arm under the halved schedule gives 60, 59 and 49, against 72, 71 and 72 at the constant rate. Screen and defect disagree in both directions (on K_4 under the matched cosine, 0 of 24 pass the screen, at a median residual of 1.6×10^{-6}, while 24 of 24 have a defect below 10^{-3}).

Deviations and Campaign E defects. The registration did not name the comparator's span for F1d-1 or the random-word estimator for F2a and F2b; both readings are reported and only F1d-1 depends on it. Reusing Campaign E's code exposed six defects, none of which moves a registered verdict of either campaign: the -1 counter tested the angle only (the retraction in Appendix E); the structured-word evaluation applied $L_{\mathrm{train}} = 12$ to every model (never-converged models only); the cover residuals were computed on the reachable span under the screen's name, with 8 words per element instead of the registered 64 (Appendix E); an overflow flag was taken from the state rather than the logits (29 unconstrained breaks unflagged); and the closed form used in Campaign E was correct only for one- and two-letter words, the only ones it was used for.

G Proofs of Lemma 1 and Propositions 1 and 2

Throughout, V is the orbit span. Under the convention $h_t = A_{x_t} h_{t-1}$ a word $u = x_1 \cdots x_T$ acts as $A_u = A_{x_T} \cdots A_{x_1}$, so the word pwq (read p first) acts as $A_q A_w A_p$, and the empty word acts as I.

The monoid hypothesis. Lemma 1 and Proposition 1 need the relators to present G as a *monoid*: two words spell the same element iff finitely many insertions and deletions of relators connect them. This holds for the four displayed presentations. For K_4, S_3 and D_4 every generator has a positive power among the relators, so it is invertible in the presented monoid, which is therefore the group. For $\mathbb{Z}_6$, write b for the separate inverse token: $a^6 = e$ makes a^5 a two-sided inverse of a, and $ab = e$ then gives $b = a^5 ab = a^5$, so $ba = e$ holds as well. The hypothesis is not automatic for an arbitrary list of identities, but it holds for every list of §4: the K_4 list gives a a positive power and $b^2 = a^{-2}$; the S_3 and D_4 lists give $s^4 = e$ and r^3 or r^4 equal to s^{-2}; the $\mathbb{Z}_6$ list gives $ba = e$ and a a right inverse $a^2 b^3$; the relator subsets and $x^2 = y^2 = e$ contain a positive power of each generator or of a product with one. *Proof of Corollary 1.* Each generator is invertible in the presented monoid, which is therefore the group $\widehat{G}$, and Lemma 1 and Proposition 1 apply with $\widehat{G}$'s presentation; the drift bound is Proposition 1's telescoping over one period.

Lemma 1.

Proof. If $A_w|_V = I$ for every $w \in \mathcal{R}$, one move replacing pq by pwq (or back) changes the state by $A_q(A_w - I)A_p h_0 = 0$, since $A_p h_0 \in V$; words spelling the same element are connected by finitely many moves, so they reach the same state. Applying this after an arbitrary prefix v gives $A_u A_v h_0 = A_{u'} A_v h_0$ for all v, so A_u and $A_{u'}$ agree on a spanning set of V and $T_g := A_u|_V$ is well defined. The word $u_g u_h$ spells gh and acts as $A_{u_h} A_{u_g}$, so $T_{gh} = T_h T_g$, and $\rho(g) := T_{g^{-1}}$ satisfies $\rho(gh) = T_{g^{-1}} T_{h^{-1}} = \rho(g)\rho(h)$ with $\rho(e) = I$. Conversely, if the state depends only on the element, then for every word u and relator w the word uw spells the element of u, so $A_w A_u h_0 = A_u h_0$; the orbit states span V, hence $A_w|_V = I$. The argument uses neither isometry nor faithfulness; if a deterministic readout classifies every element correctly, the element states are distinct and the action is faithful. □

Proposition 1. Here the A_x are isometries of V and $\|\cdot\|$ is the operator norm on V.

Proof. One move changes the state by $A_q(A_w - I)A_p h_0$; $A_p h_0 \in V$ has norm $\|h_0\|$, $(A_w - I)|_V$ has norm at most ε, and A_q is an isometry, so the change has norm at most $\varepsilon\|h_0\|$. Telescoping over N moves gives the first claim. For the second, reduce $u = x_1 \cdots x_T$ one letter at a time from $n(e) = \varnothing$: $n(g_{t-1})x_t$ is at area at most c_G from $n(g_t)$, so u is at area at most $c_G T$ from $n(g_T)$. A state perturbation of norm δ moves each logit by at most $\|W\|\delta$, hence the margin by at most $2\|W\|\delta$; the margin at the normal form is at least $\mu > 0$, so the word is read correctly while $2\|W\| c_G T \varepsilon \|h_0\| < \mu$. □

The bound $c_G T$ on the area is a linear isoperimetric (Dehn-function) bound, available because G is finite; it is what turns a per-move defect into a horizon linear in $1/\varepsilon$.

Proposition 2.

Proof. On a rotation plane P_j the two powers are $R(n\varphi_j)$ and $R(n\varphi_j^0)$, whose difference has norm $2|\sin(n(\varphi_j - \varphi_j^0)/2)| \le \min(n\delta_j, 2)$. On a sign line the difference is 0, or $(-1)^n - 1$ for a -1 line

replaced by $+1$ when m is odd, whose magnitude is at most $\min(n\pi, 2)$. The blocks are orthogonal and invariant, so the first inequality is the triangle inequality over j. For the second, let c be the class of g_u^n and $\Delta = M^n h_0 - (M^0)^n h_0$, with block components Δ_j, $\|\Delta_j\| \le \|h_j\| \min(n\delta_j, 2)$. Then

$$\mu\big(M^n h_0, c\big) = \min_{k \neq c} \big[(w_c - w_k)^\top (M^0)^n h_0 + (b_c - b_k) + (w_c - w_k)^\top \Delta\big] \ge \mu^0 - \max_{k \neq c} \big|(w_c - w_k)^\top \Delta\big|,$$

since the bracket without its last term is the ideal orbit's margin against k, and $|(w_c - w_k)^\top \Delta| \le \sum_j \|\Pi_{P_j}(w_c - w_k)\| \, \|\Delta_j\| \le \sum_j g_j \|h_j\| \min(n\delta_j, 2)$. The ideal orbit and its target classes are m-periodic, so its margin is at least μ^0 at every n. Bounding the $\mathcal{O}$ terms by 2 and the $\mathcal{D}$ terms by $n\delta_j$ gives $\mu \ge B - nC$ with $B = \mu^0 - 2\sum_{\mathcal{O}} g_j \|h_j\|$ and $C = \sum_{\mathcal{D}} g_j \|h_j\| \delta_j$. The margin is positive for $n < B/C$, and for every n when $C = 0$ and $B > 0$; if $B \le 0$ the proposition gives no guarantee, which is not a certificate of failure. Since $g_j \le 2\|W\Pi_{P_j}\|$, the bound implies the one with the spectral norm of the readout on each block in place of g_j and the constants doubled. Correctness inside a block needs the same bound at every prefix phase: after n blocks and r further tokens the state is $A_{x_1 \cdots x_r} M^n h_0$, an isometric image of $M^n h_0$, so the argument applies with W replaced by $W A_{x_1 \cdots x_r}$ and the target $g_u^n g_{x_1 \cdots x_r}$; Campaign F takes μ^0 and g_j over all phases. □

H THE RELATION PENALTY: COUNTS, PLACEBOS AND MATCHED CONTROLS

Counts. Eq. (1) runs on the orthogonal architecture with $\beta \in \{10^{-3}, 10^{-1}, 1\}$, on all four groups, both widths and both training lengths, 24 seeds each (48 configurations, 1,152 models; Campaign A): 650 of 1,152 models pass, and 659 of 1,152 on retraining from scratch in Campaign B (Table 6), after none in 192 matched controls and none in 57,920 ordinary models. Every model converges under both definitions. A pooled fraction such as $650/1{,}152$ averages over cells that range from $0/24$ to $24/24$, so it describes the design and is not an estimate of a single rate; K_4's cell medians lie three to five orders of magnitude below the threshold while S_3's lie on it, and S_3 is the only group that gets worse when β increases from 0.1 to 1, a rise that is observed, not derived (Appendix K). Because its cells sit on the threshold, the S_3 pass rate at $\beta = 0.1$ is the one number that differs between tables by cell composition: 0.50 at width 24 and $L_{\text{train}} = 24$ over 24 seeds (Table 9, Campaigns A and C), 0.26 over the 96 models of Table 22, and 0.17 and 0.08 per width in Campaign C's 12-seed cells (Table 21). Two measurements change besides the pass rate (Figure 7): the restricted relator defect falls from about 1.3 to about 10^{-4}, and the reachable span collapses from the whole of $\mathbb{R}^H$ to exactly $|G|$ dimensions, 4 for K_4, 6 for $\mathbb{Z}_6$ and 8 for D_4, the size of the regular representation; numerical rank and small defects alone do not certify its type. S_3 is the exception on both counts, with a defect of 0.42 and a span of $8 = |G| + 2$ dimensions; the two extra dimensions are the trapped plane of Appendix K.

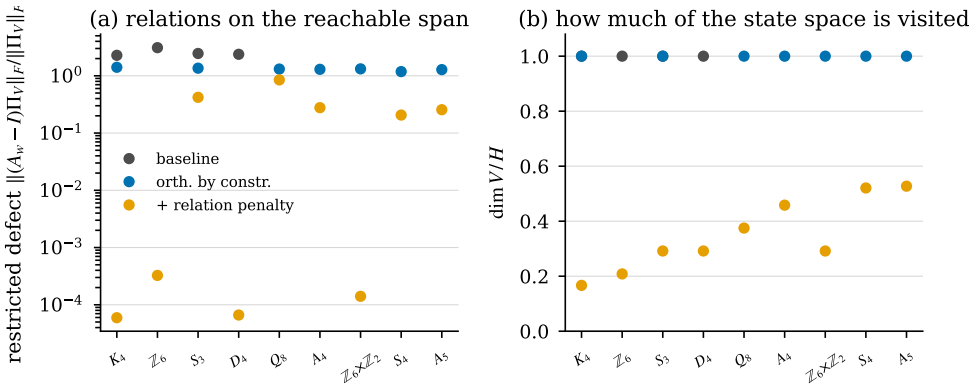


Figure 7: Measured on the models' own parameters (2,088 stored checkpoints). **(a)** Relator defect restricted to the reachable span V. **(b)** Fraction of the state space that V occupies. The relation penalty lowers the defect by three to four orders of magnitude and shrinks the span to about $|G|$ dimensions, except on the groups of Appendix J.

Dose–response. The response to β rises in every group from 10^{-3} to 0.1 (Figure 8); only on S_3, whose cells sit on the threshold, are the pass counts sensitive to τ (606 to 735 of 2,112 across the calibration window). The visit-weighted defect falls from 0.25 to 10^{-4}, and the sampled span collapses to $|G|$ dimensions, the size of the regular representation.

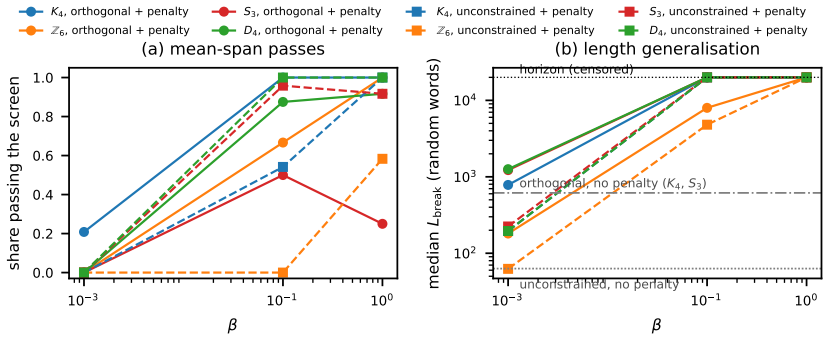


Figure 8: Dose–response, with and without the orthogonal parametrisation (width 24, $L_{\text{train}} = 24$, 24 seeds per point, constant learning rate). **(a)** Share passing the screen. **(b)** Median L_{break} under random words; points on the dotted line are censored at 20,000. Orthogonal arm: Campaigns A and C; unconstrained arm: Campaign E.

The screen against the defect on V. Table 17 sets the mean-span screen ($1 - \eta^2_{\text{sub}} < 10^{-6}$) beside a criterion on the reachable span: a restricted relator defect below 10^{-3} on V, the tolerance used for "words met" in Table 1. On K_4 and D_4 the two agree to one model in 384. On $\mathbb{Z}_6$ they disagree on 18 of 192 models, each within a factor of 3.5 of one of the two thresholds (screen-only models have defects of 1.1×10^{-3} to 3.2×10^{-3}; defect-only models have screen residuals of 1.0×10^{-6} to 3.4×10^{-6}). On S_3 the screen passes 35 models and the defect 13: the 22 screen-only models all have $\dim V = 8$ and a defect of 0.42101 to five digits, one trapped plane each (Appendix K) lying in V but outside the mean span, the blind spot of §2; the 13 that pass both have $\dim V = 6$. On the further groups the screen was not logged in this corpus, and the defect gives $\mathbb{Z}_6 \times \mathbb{Z}_2$ 20 of 24 and A_4 2 of 24, as the screen does in Table 22. Computed by `tools/v_pass_counts.py` from the stored per-model records.

Table 17: Relation penalty, orthogonal arm: mean-span screen passes against a restricted relator defect below 10^{-3} on V. Retrained corpus (Campaigns B and C; both widths and training lengths, 24 seeds) for the four original groups; Campaign D at width 24 and 48 for the further groups, where the screen was not logged in these records.

G	β	n	screen	defect on $V < 10^{-3}$	screen only	V only
K_4	0.1	96	96	96	0	0
K_4	1	96	96	96	0	0
D_4	0.1	96	83	83	0	0
D_4	1	96	88	87	1	0
$\mathbb{Z}_6$	0.1	96	88	95	1	8
$\mathbb{Z}_6$	1	96	96	87	9	0
S_3	0.1	96	25	6	19	0
S_3	1	96	10	7	3	0
$\mathbb{Z}_6 \times \mathbb{Z}_2$	0.1	24	—	20	—	—
A_4	0.1	24	—	2	—	—
Q_8	0.1	24	—	0	—	—
S_4	0.1	24	—	0	—	—
A_5	0.1	24	—	0	—	—

Placebo penalties. Under exact orthogonality, Eq. (1) amounts to "maximise $\operatorname{tr} A_w$", so one must rule out that any pressure towards the identity would do the same. Campaign B trains four alternatives with the same β, architecture and seeds, and Campaign D retrains the three informative ones with stored checkpoints so that they can be evaluated to 20,000 steps (Table 18). Penalising the *generators themselves*, which is generic shrinkage towards I, is actively harmful: almost no model converges, accuracy collapses beyond length 20, and none passes. Penalising *redundant words*, identities of G that are not the defining relators, is the informative case: those models mostly converge and extrapolate under random words as far as models trained on the true relators (at 20,000 steps a log-rank test does not separate the two arms, $\chi^2 = 1.44$, $p = 0.23$, which does not establish equal survival), yet not one of 576 passes across the two campaigns, their defect on the *defining* relators stays as large as in an unpenalised model, and their η^2_{sub} stays near 0.997; which group they represent

instead is the subject of §4 and Appendix E. The remaining two placebos, another group's relators and random words, mostly destroy training, since forcing a non-identity word to act as the identity contradicts the task; the random-word arm is additionally confounded, because in groups this small several of the drawn words happen to be true relators. We report it and do not rely on it.

Table 18: Placebo penalties on the orthogonal architecture at $\beta = 0.1$, four groups, 3,000 steps. Rows (i), (iv), (v) are from Campaign D (192 models per row; 96 for (v), width 24 only), evaluated to 20,000 steps in Campaign D's own evaluation (Campaign E's re-measurement of the same checkpoints, with 32 random sequences and per group, is Table 7: 314 and 135 for the two controls and 14,423 for the $\mathbb{Z}_6$ cell of row (i)); rows (ii), (iii) are from Campaign B (384 models per row), evaluated to 2,000. "Defect on $\mathcal{R}$" is the parameter-space defect on the group's *defining* relators, whatever words were penalised, and is not the restricted defect of Table 1.

penalty words	converged	pass	defect on $\mathcal{R}$	median L_{break}	censored
none (orthogonal control, D_4 and $\mathbb{Z}_6$)	1.00	0.00	8–16	146–319	0.00
(i) the group's defining relators	1.00	**0.78**	2–8 $\times 10^{-4}$ (S_3: 1.8–3.2)	$\geq$20,000 ($\mathbb{Z}_6$: 14,412)	0.81
(iv) redundant words (four different groups; §4, Table 1)	0.91	0.00	6–17	$\geq$20,000	0.86
(v) generators alone (shrinkage to I)	0.19	0.00	0.3–2.6	19–22	0.00
(ii) another group's relators	0.10	0.00	—	19–20	0.00
(iii) random words *(confounded, see text)*	0.22	0.13	—	19–25	—

Length generalisation against matched controls. The fair comparison holds architecture, width and training length fixed and changes only the penalty. Campaign C evaluates the retrained models to 20,000 steps (Table 19): the eight control cells on S_3 and K_4 break at medians of 60 to 840 steps with no censoring, and every one of the 384 matched penalised models survives to the horizon, a gain of at least 24- to 49-fold at $L_{\text{train}} = 24$ and 130- to 340-fold at $L_{\text{train}} = 12$. Campaign D supplies the D_4 and $\mathbb{Z}_6$ controls, which break at medians of 319 and 146 (Table 18); penalised D_4 models all survive to 20,000, and penalised $\mathbb{Z}_6$ models give the first treated breaking length that is a number rather than a bound, a median of 14,412 steps (interquartile range 10,773–18,710). For every other group the gain is a lower bound: an evaluation at 200,000 steps was too expensive for our setup, since the recurrence is stepped one timestep at a time in float64.

Table 19: Matched comparison at $L_{\text{eval}} = 20{,}000$ (Campaign C). Every penalised model is censored; the log-rank χ^2 is saturated by complete separation.

			hard-orthogonal control			+ relation penalty			
G	w	L_{train}	n	median L_{break}	cens	n	median L_{break}	cens	χ^2
K_4	24	12	24	59.5	0.00	48	$\geq$20,000	1.00	103.9
K_4	24	24	24	840	0.00	48	$\geq$20,000	1.00	104.2
K_4	48	12	24	123.5	0.00	48	$\geq$20,000	1.00	103.9
K_4	48	24	24	832	0.00	48	$\geq$20,000	1.00	104.2
S_3	24	12	24	141.5	0.00	48	$\geq$20,000	1.00	104.2
S_3	24	24	24	430.5	0.00	48	$\geq$20,000	1.00	104.2
S_3	48	12	24	151.5	0.00	48	$\geq$20,000	1.00	104.2
S_3	48	24	24	406.5	0.00	48	$\geq$20,000	1.00	104.2

Numbers that depend on the population. Several quantities are reported for more than one population or evaluation run; Table 20 lists them side by side.

Table 20: Quantities reported for more than one population or evaluation run.

quantity	value	population or run
S_3 screen pass rate, orthogonal + penalty, $\beta = 0.1$	0.50	width 24, $L_{\text{train}} = 24$, 24 seeds (Table 9)
	0.26	96 models, both widths and lengths (Table 22)
	0.17 / 0.08	Campaign C, 12 seeds per width (Table 21)
	6 of 96	same models as 0.26, by the defect on V (Table 17)
D_4 / $\mathbb{Z}_6$ orthogonal controls, random words	319 / 146	Campaign D's evaluation (Table 18)
	314 / 135	Campaign E's re-measurement, 32 sequences (Table 7)
$\mathbb{Z}_6$ defining relators, random words	14,412 / 14,423	Campaign D / Campaign E evaluation of the same checkpoints
K_4 redundant words, $(ab)^n$	822 / 954	all 48 models / 42 converged models
penalty passes, $\beta \in \{0.1, 1\}$, four groups	582 / 557 of 768	screen / defect on V (Table 17)

I THE PENALTY UNDER CONSTANT AND DECAYING LEARNING RATES

Under the constant rate of 3×10^{-3} the relator defect of the penalised orthogonal arm reaches a minimum between 2,500 and 4,000 steps, rises by about two orders of magnitude, and settles on a plateau, while accuracy stays perfect from step 500 onwards; the pass fraction at 10,000 steps falls accordingly on D_4, $\mathbb{Z}_6$ and $\mathbb{Z}_6 \times \mathbb{Z}_2$ (Campaign C, Table 21), and configurations that survive to 20,000 evaluation steps when stopped at 3,000 break between 1,400 and 9,100 when trained to 10,000. Keeping, for each model, the checkpoint with the smallest *training* relator defect recovers the 3,000-step pass rate in nine cells out of ten and exceeds it in the tenth. Campaign D reran the affected groups with the same seeds under a cosine decay of the learning rate to zero and under a constant rate three times smaller. Under the cosine schedule the pass rate at 10,000 steps equals the rate at 3,000 and the defect keeps falling, to 300 to 1,000 times below the constant-rate plateau in every group except S_3, whose trapped plateau is identical to four digits under both schedules (ratio 0.99996), consistent with the penalty minimum of Appendix K; two tested schedules do not establish invariance to every schedule or optimiser. Under the constant 10^{-3} the plateau is 0.24 to 0.32 times the one at 3×10^{-3}, proportional to the learning rate. With the schedule, median $1 - \eta^2_{\mathrm{sub}}$ reaches 10^{-11}, and all penalised K_4, D_4, S_3 and $\mathbb{Z}_6 \times \mathbb{Z}_2$ models trained to 10,000 steps survive to the 20,000-step horizon. The training-time optimum is therefore an artefact of a constant step size, and the penalty results of §6 should be read as "relation penalty with a decaying learning rate, or stopped on the training defect".

Why a short cosine hurt the unconstrained arm. Campaign E's 3,000-step cosine decay to zero halves the integrated learning rate. Campaign F matched the integral (peak 6×10^{-3}): the unconstrained arm at $\beta = 0.1$ then passes the screen in 40 of 96 models and has a defect below 10^{-3} on V in 65, against 0 and 23 under the halved schedule and 60 and 70 at the constant rate; the orthogonal arm under the halved schedule gives 60 and 59, against 72 and 71 at the constant rate (Appendix F). The integrated rate accounts for most of the difference between the arms that Campaign E could not explain.

Table 21: Fraction passing under a constant learning rate of 3×10^{-3} (Campaign C) at 3,000 and 10,000 steps and at the checkpoint with the smallest *training* relator defect, and at 10,000 steps under a cosine schedule and under a constant 10^{-3} (Campaign D; same seeds, 12 per cell). Width 24 / width 48; every model converges under both definitions. †Stopped early by a time limit (step 5,000 and step 6,000), not cited in the text.

	constant 3×10^{-3}			cosine to 0	constant 10^{-3}	
G	@3,000	@10,000	@min defect	@10,000	@10,000 (w24)	defect @10,000: const. / cosine (w24)
K_4	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / —	—	1.8×10^{-2} / 3.0×10^{-5}
S_3	0.17 / 0.08	0.17 / 0.17	0.25 / 0.08	0.33 / —	—	1.842 / 1.842
D_4	0.92 / 0.83	0.67 / 0.50	0.92 / 0.83	**0.92 / 0.75**	0.92	2.3×10^{-2} / 3.6×10^{-5}
$\mathbb{Z}_6$	1.00 / 1.00	0.58 / 0.75	1.00 / 1.00	**1.00** / 1.00†	1.00	2.3×10^{-2} / 7.3×10^{-5}
$\mathbb{Z}_6 \times \mathbb{Z}_2$	0.83 / 0.92	0.33 / 0.33†	0.83 / 0.92	**0.83** / —	0.92	3.9×10^{-2} / 3.6×10^{-5}

J THE FURTHER GROUPS, AND THE TRAPS

Table 22: Nine groups under the relation penalty at $\beta = 0.1$, orthogonal arm (Campaigns B and C; medians at width 24, 64 for A_5). The last column is the un-penalised orthogonal arm's η^2_{sub} on the same group; it passes 0 everywhere. Passes concern the mean span. Trapped costs are computed in closed form, with no measured input, in Table 23.

G	$\lvert G\rvert$	$d_{\min}$	n	structure	**pass**	$1 - \eta^2_{\mathrm{sub}}$	restr. defect	orth. η^2_{sub}
K_4	4	2	96	abelian	**1.00**	$< 10^{-8}$	6×10^{-5}	0.999
$\mathbb{Z}_6$	6	2	96	abelian	**0.92**	$< 10^{-7}$	3×10^{-4}	0.998
S_3	6	2	96	non-abelian	0.26	6×10^{-6}	0.42 (trapped)	0.999
D_4	8	2	96	non-abelian	**0.86**	$< 10^{-8}$	7×10^{-5}	0.998
Q_8	8	4	24	non-abelian, quaternionic	0.00	8×10^{-4}	0.80 (trapped)	0.9990
$\mathbb{Z}_6 \times \mathbb{Z}_2$	12	3	24	abelian	**0.83**	$< 10^{-6}$	2×10^{-4}	0.9956
A_4	12	3	24	non-abelian	0.08	$< 10^{-4}$	0.28 (trapped)	0.9965
S_4	24	3	24	non-abelian	0.00	4×10^{-3}	0.22 (trapped)	0.9903
A_5	60	3	24	non-solvable	0.00	3×10^{-2}	0.27 (trapped)	0.9836

The mechanism extends; the remedy only partly. On all five further groups the baseline and orthogonal arms pass none of 240 models, with η^2_{sub} between 0.60 and 0.83 for the baseline and between 0.977 and 0.999 for the orthogonal arm. The penalty works on the abelian $\mathbb{Z}_6 \times \mathbb{Z}_2$, where 83% of models pass and the reachable span shrinks from 24 to 10 dimensions, and fails on the other four: A_4 passes 2 of 24 models, and Q_8, S_4 and A_5 none (Table 22). It is not inert there: it reduces the restricted defect by a factor of 1.6 (Q_8) to 5 (S_4) relative to the orthogonal arm, compresses the span, and every penalised model still survives to the 20,000-step horizon, which unpenalised models on these groups do not approach; but it stops three to four orders of magnitude short of a representation. Two matched pairs carry the comparison: Q_8 and D_4 have the same order and the same character degrees, and their pass rates differ by 0.86; Q_8's two-dimensional irrep is quaternionic, so its smallest faithful real representation has four dimensions, and its presentation has no generator of order 2. $\mathbb{Z}_6 \times \mathbb{Z}_2$ and A_4 have the same order and the same $d_{\min}$ and differ by 0.75; one of them is abelian. Across the nine groups the three abelian ones pass at 0.83 or better, and among the six non-abelian ones only D_4 exceeds 0.26. Width does not help: S_4 run at widths 24, 48, 96 and 192 (one to eight times $|G|$, 12 seeds each) passes no model at any width, and its relator defect *rises* from 1.95 to 4.83 while the code grows towards the regular representation.

The failure is immediate and persistent, and on A_4 and S_4 quantised. On Q_8, A_4, S_4 and A_5 the relator defect reaches its floor within 500 steps and then changes by less than 0.06% over the next 9,500: persistence over the observed interval, not a proof of permanent trapping. The floors are far from arbitrary (Figure 9). The trap statistics use Campaign D's 36 penalised models per group (A_4, Q_8, S_4: 24 at width 24 and 12 at width 48; A_5: widths 64 and 128) and, for S_3, the 288 penalised models of the retrained corpus (three β, widths 24 and 48), of which 265 are trapped; they are not the 24-model cells of Table 22. For each trapped model we take the squared relation target $\sum_w \|A_w - I\|_F^2$ and divide it by the smallest value seen in its group: the models within 2% of a whole number are S_3 265 of 265 (unit 3.3934, levels 1–4), A_4 32 of 32 (2.2782, levels 1–3), S_4 35 of 36 (1.2635, levels 1–7) and Q_8 19 of 36 (3.3841, levels 1–5); A_5 (5 of 36) is the exception, for the reason given below. The rise of S_4's defect with width is the same fact seen from another angle, since it grows from about 3 units at width 24 to about 18 at width 192.

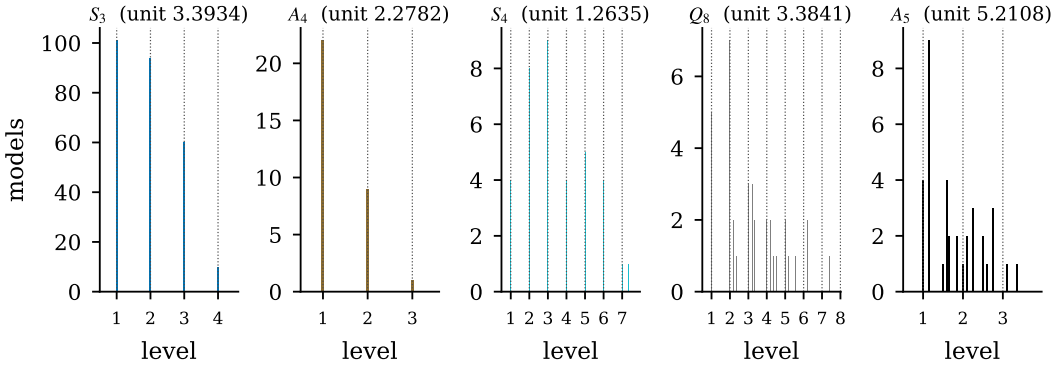


Figure 9: Defect floors come in whole-number multiples. For each trapped model under the relation penalty, the squared relation target is shown in units of the smallest value in its group; dotted lines mark the integers. The unit for S_3 is the closed-form constant of Appendix K.

The units are local minima of the penalty, computed without any data. From every penalised checkpoint we extracted the invariant 2-planes on which a generator rotates by an angle its order does not permit (more than 0.05 from every permitted angle, which excludes the arbitrary planes an eigensolver returns inside degenerate eigenspaces), and recorded on each the other generator's angle and invariance residual and each relator's contribution. For S_3 this leaves 508 planes at width 24, invariant under the other generator to a median residual of 4.5×10^{-5} to 2.4×10^{-4}, at angles 0.434561 and 2.272471, overlapping the reachable span at 0.999996; for S_4 the residuals are 10^{-4} to 3×10^{-3} at angles 0.3146 and 1.6455 with per-plane cost 1.263527; for Q_8, 78 of 200 such planes are jointly invariant, at angle pairs $(1.6309, 2.2621)$ and $(1.5107, 0.8795)$ with cost 3.384061, both images of the computed minimiser $(1.630925, 0.879475)$ under the penalty's symmetries $(x, y) \mapsto (-x, \pi - y)$ and $(\pi - x, -y)$, with angles reported in $[0, \pi]$; for A_5 the two generators' planes cost 0.2178 and 3.2509; for A_4 the residuals are 1.000 and $1.1547 = 2/\sqrt{3}$ in every plane, a structured value that signals a

larger invariant block. Independently, for a plane invariant under both generators on which they act as rotations by x and y, each relator w contributes $4(1-\cos\phi_w)$ with ϕ_w the total angle of w; we minimised the resulting function from 1,500 random starts per group, with no measured quantity as input, and collected the distinct local minima (Table 23). S_3: $\{3.393370\}$. S_4: $\{1.263527,\ 4.662340,\ 8\}$, the first at $(x,y)=(\pi-0.314564,\ \pi-1.645456)$. Q_8: $\{3.384061,\ 8\}$ at $(1.630925, 0.879475)$. A_5: a family of seven, $\{0.217769,\ 0.859651,\ 1.890877,\ 3.250893,\ 4.843831,\ 6.498124,\ 7.787606\}$, which is why its floors are sums over several constants and not multiples of one. A_4: $\{4.283953\}$ only, which is not observed; minimising instead over pairs in $SO(n)$ parametrised by the exponential map, $n=4,5$ give $\{0,\ 2.27821,\ 4.28395,\ 8.56791\}$, so A_4's trap is a four-dimensional block. The two computations agree to every printed digit. To 60 digits the S_4, Q_8 and smallest A_5 constants are $1.26352688538990\ldots$, $3.38406052281249\ldots$ and $0.21776902193908\ldots$; an integer-relation search finds no polynomial of degree up to 8 with coefficients up to 2×10^4 for any of them, so unlike S_3's constant they are characterised as minima of an explicit trigonometric function rather than by a radical expression.

Table 23: Trapped configurations. "Computed": smallest nonzero local minimum of the relator penalty restricted to the stated block, from random starts with no measured input (for A_5, two of a family of seven). "Measured": per-block cost in trained models.

G	block	restricted penalty	computed	measured	generator angles (measured)
S_3	plane	$4[3-\cos3\alpha-\cos2\theta-\cos(2\alpha+2\theta)]$	3.393370	3.3934	2.272471, 0.434561
S_4	plane	$4[3-\cos2x-\cos4y-\cos3(x+y)]$	1.263527	1.263527	0.3146, 1.6455
Q_8	plane	$4[3-\cos4x-\cos(2x+2y)-\cos(2x+4y)]$	3.384061	3.384061	1.6309, 2.2621
A_5	plane	$4[3-\cos2x-\cos5y-\cos3(x+y)]$	0.217769, 3.250893	0.2178, 3.2509	—
A_4	$SO(4)$ block	$\sum_w\|A_w-I\|_F^2$ on $SO(4)$	2.27821	2.2782	(not a plane)

K THE S_3 TRAP IN CLOSED FORM

The plateau. Define the *relation target* as $(\sum_{w\in\mathcal{R}}\|A_w-I\|_F^2)^{1/2}$ on the model's parameters. For K_4 the penalty drives it to machine zero. For S_3 it stops at a plateau that three orders of magnitude of β do not move, quantised by the number k of complex eigenvalues of A_s: about 0.001 when $k=0$, 1.8421 when $k=2$, and 2.6051 when $k=4$, a ratio of $\sqrt{2}$ to five significant figures. A complex pair means that A_s acts on some plane P as a rotation $R(\theta)$ with $\theta\notin\{0,\pi\}$, so $A_s^2\neq I$ there. Suppose, and this is an assumption that is not forced, that P is invariant under both generators and that A_r rotates it too. Rotations of a plane commute, so if the relations held on P the action there would factor through the abelianisation of S_3, which is $\mathbb{Z}_2$, and the plane could not carry the faithful two-dimensional irrep that the rest of the state is built from. The plane is stuck: three relators compete for two angles.

Derivation. On a 2-plane V invariant under both generators with $A_s|_V=R(\theta)$, $A_r|_V=R(\alpha)$, and using $\|R(\phi)-I\|_F^2=4(1-\cos\phi)$, the three relators r^3, s^2, $(rs)^2$ contribute

$$P(\alpha,\theta)=4\big[\,3-\cos3\alpha-\cos2\theta-\cos(2\alpha+2\theta)\,\big].$$

$\partial P/\partial\theta=0$ gives $\sin2\theta+\sin(2\alpha+2\theta)=0$, whose solution set is $\alpha=k\pi-2\theta$ together with $\cos\alpha=0$; the latter, combined with $\partial P/\partial\alpha=0$, requires $|\sin2\theta|=3/2$ and is empty. On the even-k branch $\alpha=-2\theta$, $P=4[3-\cos6\theta-2\cos2\theta]$: its points with $\sin2\theta=0$ are the zero-penalty configurations ($2\theta\equiv0$, $\alpha\equiv0$, the trivial and sign actions on the plane) and a point with $P=24$, and its other critical points have $\sin^2 2\theta=11/12$ and $P\in\{11.2302,\ 12.7698\}$, far above the observed plateau. Substituting the odd-k branch $\alpha=\pi-2\theta$ into $\partial P/\partial\alpha=0$ and expanding $\sin3u=3\sin u-4\sin^3u$ yields $12\sin^3(2\theta)=7\sin(2\theta)$, so either $\sin2\theta=0$, which on this branch gives $P\in\{8,16\}$ (both values reappear in the census below), or

$$\sin^2(2\theta)=\tfrac{7}{12},\qquad \cos2\theta=\tfrac{\sqrt{15}}{6},\qquad \theta^\star=\tfrac12\arccos\tfrac{\sqrt{15}}{6}=0.434561101504.$$

Back-substitution gives $\|A_{r^3}-I\|_F^2=4-8\sqrt{15}/9=0.557348136705$ and $\|A_{s^2}-I\|_F^2=\|A_{(rs)^2}-I\|_F^2=4-2\sqrt{15}/3=1.418011102528$, summing to $12-20\sqrt{15}/9=3.393370341761$ per trapped plane, with $\cos6\theta=4\cos^32\theta-3\cos2\theta=-\frac{2}{9}\sqrt{15}$. The per-plane contribution is $c:=\big(\frac12\sum_w\|A_w-I\|_F^2\big)^{1/2}$, so the relation targets are $c\sqrt{2}=1.8421$ for one trapped plane ($k=2$) and $2c=2.6051$ for two ($k=4$). Squaring twice eliminates the radical, $(c^2-6)^2=500/27$, i.e.

$27c^4 - 324c^2 + 472 = 0$, irreducible over $\mathbb{Q}$ and hence the minimal polynomial of c. At $(\alpha^\star, \theta^\star)$, where $2\alpha + 2\theta \equiv -2\theta$ and $\cos 3\alpha = -\cos 6\theta$, the Hessian of P is $\frac{8\sqrt{15}}{3}\left(\begin{smallmatrix} 4 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, with eigenvalues $\frac{8\sqrt{15}}{3}(3 \pm \sqrt{2}) = 16.378$ and 45.590: a strict local minimum of the restricted penalty. Because it is an exact critical point of the penalty, $\nabla(\beta P) = 0$ for every $\beta > 0$: rescaling the penalty keeps its stationary configurations and scales their curvature, which is why the plateau ignores the penalty weight. The joint task-plus-penalty objective need not share them, so the rise in trapping from $\beta = 0.1$ to 1 is observed, not derived. Zero-penalty configurations exist, so the minimum is an obstruction of the landscape, not of expressivity. Table 24 sets the closed form against measurement; θ was not a fitting target, and the angle we had guessed beforehand, $\pi/2$, occurs in none of the 580 measured eigenvalues.

Table 24: Closed form against measurement. θ is not a fitting target.

quantity	closed form	derived	measured	rel. error
$\|A_{r^3} - I\|_F^2$	$4 - 8\sqrt{15}/9$	0.557348136705	0.5573477778	6.4e-07
$\|A_{s^2} - I\|_F^2$	$4 - 2\sqrt{15}/3$	1.418011102528	1.4180088952	1.6e-06
$\|A_{(rs)^2} - I\|_F^2$	$4 - 2\sqrt{15}/3$	1.418011102528	1.4180150333	2.8e-06
sum	$12 - 20\sqrt{15}/9$	3.393370341761	3.3933717062	4.0e-07
c	$\sqrt{6 - 10\sqrt{15}/9}$	1.302568681829	1.3025689437	**2.0e-07**
θ	$\frac{1}{2}\arccos(\sqrt{15}/6)$	0.434561101504	0.4345609627	**0.04** σ

The explicit construction, and confirmations. With $\theta = \theta^\star$ and $\alpha = \pi - 2\theta^\star$, set $A_r = R(2\pi/3) \oplus R(\alpha) \oplus I_2$ and $A_s = \mathrm{diag}(1, -1) \oplus R(\theta) \oplus \mathrm{diag}(-1, 1)$. Both have $\det = +1$, so both lie in $SO(6)$ and are reachable by the exponential parametrisation; A_s has exactly one complex pair and an even-dimensional -1 eigenspace, as observed in 100% of models. The first summand carries the faithful 2-dimensional irrep exactly, the third a further reflection pair, and the middle summand is the trapped plane. The construction reproduces all three closed-form defects to twelve digits (the measured values agree with them to the relative errors of Table 24), and two orthogonal trapped planes in an $SO(8)$ analogue double the sum exactly (ratio 2.000000000000), which is what makes $c\sqrt{k}$ a consequence of Frobenius additivity over orthogonal blocks rather than a fitted scaling law. The criticality is not an artefact of the two-angle reduction: the full 30-dimensional gradient of the penalty over $\mathfrak{so}(6)^2$ at this point is 5×10^{-10}, with no negative Hessian directions, the flat ones being the simultaneous-conjugation gauge orbit, and the Lanczos fixture of Campaign E on the same construction gives gradient 4.6×10^{-14} and smallest non-gauge Hessian eigenvalue $+19.36$ in the coordinates of the training parametrisation (Appendix E), which are not the angle coordinates of the 2×2 Hessian above; the two computations agree that the point is a numerically supported local minimum off the gauge orbit at this $SO(6)$ construction, which does not prove the same Hessian structure at every width or embedding; the eigenvalues are not to be compared across coordinates. Three quantities follow from $(\alpha^\star, \theta^\star)$ without being fitted: the equality of the s^2 and $(rs)^2$ defects; $3\alpha \equiv \pi - 6\theta = 0.5342260$ against 0.5342259 measured; and $\theta^\star$ itself, 0.04σ from the measured $0.4345609627 \pm 3.84 \times 10^{-6}$ over 580 eigenvalues. A critical-point census of 4,000 solver runs from random starts, using no measured value, returns $\{0,\ 3.393370,\ 8,\ 11.230200,\ 12.769800,\ 16,\ 20.606630,\ 24\}$, in which 3.393370 is the unique nonzero value matching the data and the smallest nonzero one; L-BFGS on the relator penalty alone, with no task loss, no β and no data, lands on $c = 1.30256862$ and 1.30256871 from random initialisations. The derivation used Campaign-A models; in Campaign C the squared relation target of all 265 trapped S_3 models from the retrained corpus is a whole-number multiple of $3.3934 = 2c^2$ at levels 1 to 4, and on 244 trapped planes of Campaign D the measured angles are 0.434561 for A_s and 2.272471 for A_r, against predictions of $\theta^\star = 0.434561$ and $\pi - 2\theta^\star = 2.272470$, each plane invariant under the other generator to a residual of about 5×10^{-5}, which supports the two assumptions of the derivation on an independent corpus.