ROTOR RECURRENCES: WHAT COMPACT RECURRENT STATES CAN PREDICT

Anonymous authors
Paper under double-blind review

ABSTRACT

Compact recurrent states can preserve order-sensitive computations, but what a state can store need not be what a continuous decoder can recover.

For pure rotor transitions associated with rotations in three or four dimensions, we prove a sharp prediction limit for any fixed finite number of heads. Under i.i.d. inputs from a fixed distribution whose target labels generate the group, no fixed continuous strictly positive decoder can asymptotically achieve expected prefix-averaged log loss below $\log 60$ when predicting permutations of five objects. The bound is sharp: a parity-tracking recurrence approaches this floor with a decoder that assigns almost all probability mass uniformly within the correct parity class. The same rotor family nevertheless permits arbitrarily small loss when tracking even permutations alone. A general compact-group result characterizes the hidden ambiguity underlying these limits.

We then study related affine rotor models with additive writes and decay, which lie outside the theorem. A single rotor layer learns the even-permutation task and generalizes to sequences four times the training length. Across implemented parameterizations of the same ideal rotation family, median sustained solve times among successful runs span roughly a factor of five. These experiments measure task learning without identifying the learned internal representation.

Together, the results distinguish what recurrent dynamics can represent, what continuous decoding can recover, and what gradient training discovers.

1 INTRODUCTION

A recurrent model must summarize an arbitrarily long sequence in a fixed-size state. The central challenge is therefore not only how much information the state can store, but what information about the sequence remains predictably recoverable from that state as the sequence grows. Non-commutative state transitions provide one promising mechanism: unlike commuting updates, their products depend on input order, allowing a fixed-size state to preserve computations that cannot be reduced to an order-insensitive summary. Rotations give the simplest picture: turning a book about two different axes generally gives a different result when the two turns are reversed.

Yet representability and prediction are not the same problem. A recurrent state may encode distinctions between histories while a neural readout cannot reliably recover those distinctions. In particular, standard neural decoders are continuous, so nearby states must produce nearby probability distributions. This raises a more quantitative question than whether exact state tracking is possible: when exact prediction is inaccessible to a continuous decoder, how much predictive information remains, and what is the best achievable asymptotic loss?

Recent work has established expressivity limits and exact-tracking constructions for restricted recurrent transition families, including structured linear and non-commutative state-space models (Merrill et al., 2024; Shakerinava et al., 2026; Siems et al., 2025; Walker et al., 2025; Terzic et al., 2025; Siems et al., 2026). Other analyses characterize quotient-like or partial solutions when full state tracking is not obtained (Zhang & Li, 2026). These results clarify when a transition family can or cannot represent a target computation, but they do not characterize optimal continuous-decoding loss for the compact non-commutative transition families studied here. The missing quantity is the residual ambiguity induced by the transition representation.

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.

Table 1: Three properties of a compact recurrent state and where this paper tests each. Algebraic representability need not give continuous prediction; available constructions need not be learned under the tested protocols.

property	question	evidence here
representability	can the transition family carry the computation?	A_5, S_4: finite lifts in $\mathrm{Spin}(3)$; S_5: exact tracking with a discontinuous decoder (Appendix A, Remark 1)
predictive limit	what can a fixed continuous decoder recover from the state as length grows?	$\log 60$ floor on S_5 for pure rotor state, sharp; arbitrarily small loss on A_5 (Theorem 1)
learnability	does gradient training solve a task the family can represent?	A_5 learned and extrapolated; S_4 and S_5 not reliably; up to fivefold gap between parameterizations (§4)

We characterize this ambiguity through a hidden target fibre. Consider a compact transition group G, a finite target group H, and i.i.d. inputs whose labels generate H. In the closure of the joint walk of accumulated transitions and targets, several target elements can accompany the same transition; finite words may still be decoded exactly, but the ambiguity appears in the limit. We collect these indistinguishable elements into a subgroup J. Our main result shows that J quantifies the residual predictive ambiguity: when the full accumulated transition is observed, the optimal asymptotic prefix-averaged log loss over fixed continuous positive decoders has infimum $\log |J|$; for continuous state observations, $\log |J|$ remains a lower bound, and the infimum equals $\log |J|$ when the corresponding quotient is continuously recoverable from the state. Thus the result does not merely rule out exact prediction; it quantifies the best continuous prediction that remains after exact prediction becomes inaccessible.

We then specialize this principle to compact rotor states. For pure finite-head $\mathrm{Spin}(3)$ and $\mathrm{Spin}(4)$ transitions tracking permutations in S_5, the rotor family's subgroup structure (no closed subgroup of finitely many $\mathrm{Spin}(3)$ or $\mathrm{Spin}(4)$ left actions has a continuous S_5 quotient) forces the hidden fibre to contain A_5, and hence $|J| \geq 60$. The resulting lower bound is therefore $\log 60$, and it is sharp: a continuous decoder that retains parity, spreads almost all of its mass uniformly over the 60 permutations in the recorded parity class and leaves a small remainder elsewhere approaches $\log 60$. In contrast, the same rotor family admits arbitrarily small loss on A_5, whose elements can be lifted into a finite subgroup of $\mathrm{Spin}(3)$. The S_5/A_5 contrast isolates the predictive obstruction as a continuous-decoding limitation rather than a generic inability of rotors to represent non-commutative computations.

Finally, we study what gradient training discovers when additional memory mechanisms are restored. The empirical model uses affine rotor updates with additive writes and decay, and therefore lies outside the assumptions of the predictive-limit theorem. We use these models to separate representability from learnability rather than to claim that the theorem proves an optimization failure. On A_5, a single $\mathrm{Cl}(3,0)$ rotor layer reaches accuracy 1.000 at lengths 128, 256, and 512 after training at length 128, demonstrating fourfold length extrapolation. Different parameterizations of the same ideal $\mathrm{Spin}(3)$ family also exhibit substantially different solve-time distributions. On S_4 the tested configurations exhibit partial solutions, and on S_5 parity-dominated behavior, rather than reliable full-product tracking. These experiments provide a complementary view of how optimization interacts with the structural predictive limits.

Our contributions are therefore threefold (Table 1). First, we give a general compact-group hidden-fibre characterization of optimal continuous-decoding loss. Second, we derive the sharp $\log 60$ asymptotic prediction floor for pure finite-head rotor states on S_5, together with the contrasting arbitrarily small-loss result on A_5. Third, we study learning in related affine rotor models, establishing successful A_5 extrapolation while documenting optimization sensitivity and partial solutions on harder targets.

2 ROTOR RECURRENT STATE

We study a state vector $S_t \in V$ in a real vector space V equipped with a bilinear product (an algebra).

2.1 PURE TRANSITION DYNAMICS

For the theoretical analysis, we isolate the predictive capacity of the transitions themselves. We define the pure transition recurrence:

$$S_t = R_t S_{t-1}, \qquad R_t \in G \tag{1}$$

where G is a compact transition group acting on the state by left multiplication. This transition-only mathematical object is what Theorem 1 characterizes.

2.2 AFFINE ROTOR MODEL USED FOR LEARNING

For the empirical studies, we introduce additive writes $U_t \in V$ and a scalar decay $\lambda_t \in (0, 1)$:

$$S_t = \lambda_t R_t S_{t-1} + U_t. \tag{2}$$

The affine model adds memory mechanisms that place it outside the predictive-limit theorem; we study it separately as a model of learnability. Its pairs compose associatively, $(A_2, B_2) \circ (A_1, B_1) = (A_2 A_1,\ A_2 B_1 + B_2)$, permitting a parallel prefix scan (Blelloch, 1990). Per head, with $q_t, k_t, v_t \in \mathbb{R}^n$ from linear projections,

$$U_t = k_t\, v_t = k_t \cdot v_t + k_t \wedge v_t, \qquad T_t = \lambda_t R_t, \quad \lambda_t = \sigma(w_\lambda^\top x_t),\ R_t \in \mathrm{Spin}(n), \tag{3}$$

$$S_t = T_t S_{t-1} + U_t, \qquad y_t = \langle q_t S_t \rangle_1, \tag{4}$$

followed by RMSNorm, a GLA-style output gate (Yang et al., 2023), an output projection and a SwiGLU block. Two write variants are implemented: the outer write $U_t = k_t v_t$ above, used in the language-model runs (Appendix C), and the convex write $U_t = (1 - \lambda_t)\, k_t v_t$, used in the group-tracking protocol (Appendix B.1) and compared with the outer write in Appendix B.6. The implementation allocates 2^n state scalars per head and scans all of them. In the $\mathrm{Cl}(3, 0)$ group-task runs with a learned initial state all eight coordinates can be populated; under even-rotor updates the odd coordinates evolve independently and do not reach the grade-one readout, and reported payloads count all eight. Left multiplication is linear on the state, so the recurrence is first-order linear; three independent scans agree with an unrolled product to 10^{-12} in double precision.

2.3 WHY NON-COMMUTATIVITY IS AVAILABLE, AND WHY THAT IS NOT ENOUGH

Tracking a running product whose order matters needs transitions that do not commute: applying A then B must differ from B then A. For pure transition-only updates, scalar or fixed-axis phase transport lies in a commuting subgroup, so products of such transitions do not depend on their order (though in affine recurrences, additive writes can still preserve sequence order). For $n = 3$ the even subalgebra $\mathrm{Cl}^+(3, 0) \cong \mathbb{H}$ is the quaternions, a rotor is a unit quaternion, and $R_3 R_2 R_1 \neq R_1 R_2 R_3$ in general. Formally, the transition acts on the state by left multiplication, $\rho_L(R)S = RS$, so the transition group is $\rho_L(\mathrm{Spin}(n))$ rather than the induced $\mathrm{SO}(n)$: the state distinguishes R from $-R$. For $n = 2$ the even subalgebra is $\mathbb{C}$ and the recurrence is gated linear attention over $\mathbb{C}$, sharing complex-phase transport with Selective RoPE (Movahedi et al., 2025) and Mamba-3 (Lahoti et al., 2026); we study $n = 3$.

2.4 PARAMETERIZATIONS OF THE UNIT ROTOR GROUP

We use three parameterizations of $\mathrm{Spin}(3)$: the Euler product $R = \prod_p \exp(\theta_p B_p)$ over bivector planes, the exponential map $R = \exp(B)$, and normalization $R = q/\|q\|$. The normalized implementation uses $q/\sqrt{\|q\|^2 + 10^{-6}}$, a scaled rotor whose extra contraction is absorbable into λ_t; the paired intervention of §4.4 uses unit-direction normalization with a denominator floor. Exact group statements refer to ideal unit rotors. To reach $\mathrm{Pin}(n)$ the transition may be an odd versor $u_t R_t$ selected per head by a hard gate with a straight-through estimator, since $\mathrm{Pin}(n)$ has two components. The parameterizations cover the same ideal group; §4.2 measures whether they optimize alike.

3 PREDICTIVE LIMITS OF COMPACT ROTOR STATE

Removing writes and decay isolates the information carried by the accumulated transition. The decoder class throughout includes ordinary neural readouts: fixed continuous finite-valued logits followed by softmax are continuous and strictly positive on the compact state orbit. The decoder may be trained; it is held fixed as the evaluation horizon grows.

3.1 HIDDEN-FIBRE PRINCIPLE

Fix a compact transition group G, maps $T : \Sigma \to G$ and $\pi : \Sigma \to H$ into a finite target group H, and i.i.d. inputs with positive mass on every symbol whose labels generate H. The token pairs $(T(x), \pi(x))$ generate a closed subgroup K of $G \times H$, on which the accumulated transition g_t and the target h_t evolve together. Let $P = p_G(K)$ be the closure of the reachable transitions. The target labels paired with the identity of G, those the accumulated transition cannot distinguish, form a normal subgroup $J \trianglelefteq H$, the hidden target fibre, and the induced continuous homomorphism $P \to H/J$ maps g_t to the coset $h_t J$. Proposition 1 (Appendix A.2) shows that, over fixed continuous strictly positive decoders of the full transition g_t, the infimum of the asymptotic prefix-averaged log loss is exactly $\log |J|$. For an observed state gs_0 the lower bound still holds; equality is guaranteed if the coset remains continuously recoverable from the state. The proof has two steps: Cesàro averages of the joint walk converge to Haar measure on K (Kawada & Itô, 1940, Thm. 7), under which the target is uniform on a J-coset given g_t, and Jensen's inequality gives the floor $\log |J|$; spreading mass uniformly over the coset approaches it. The bound is asymptotic, not a finite-length information or accuracy bound, and it gives no finite-length rate or optimization guarantee. Strict positivity of the decoder only keeps the loss bounded; truncating the loss at a large constant gives the same lower bound for the $\liminf$ without it.

The fibre depends on the transition group. Abelian G forces $[H, H] \subseteq J$, since H/J is a quotient of the abelian group P, so pure compact abelian transitions have optimal loss $\log 60$ on A_5 (uniform prediction), where rotor transitions reach arbitrarily small loss (Corollary 1). On S_5 the two floors coincide, since $A_5 = [S_5, S_5]$; on $H = S_5 \times A_5$ they separate: with at least two $\mathrm{Spin}(3)$ heads or one $\mathrm{Spin}(4)$ head the rotor floor is $\log 60$, strictly below the abelian floor $\log 3600$ (Appendix A.2).

3.2 ROTOR OBSTRUCTION

For A_5, the rotor family carries the target: the binary icosahedral subgroup of $\mathrm{Spin}(3)$ maps onto A_5, giving a finite set of rotor states that tracks it exactly, so the fibre can be trivial. For S_5, continuous decoding faces a structural boundary: no closed subgroup of finitely many $\mathrm{Spin}(3)$ or $\mathrm{Spin}(4)$ left actions has S_5 as a continuous quotient (Lemma 1). This is the rotor-specific step: it forces $J \neq 1$, and since J is normal in S_5, $A_5 \subseteq J$ and $|J| \geq 60$. The theorem turns this obstruction into a prediction-loss bound for the observed state, without requiring exact tracking or a separation margin.

Theorem 1 (Asymptotic loss floor for pure rotor prediction of S_5)**.** *Let $x_1, x_2, \ldots$ be i.i.d. from a distribution with positive mass on every symbol of a finite alphabet Σ, let $\pi : \Sigma \to S_5$ have image generating S_5, and let $h_t = \pi(x_t) \cdots \pi(x_1)$. Let $G = \rho_L(\mathrm{Spin}(3))^{n_\mathrm{h}}$ or $\rho_L(\mathrm{Spin}(4))^{n_\mathrm{h}}$ for any finite n_h, let $T : \Sigma \to G$, $g_t = T(x_t) \cdots T(x_1)$ and $s_t = g_t s_0$ for a fixed initial state s_0. Let p be a fixed continuous map from the orbit $\{gs_0 : g \in G\}$ to probability vectors over S_5 with every entry strictly positive. Then*

$$\lim_{L\to\infty} \frac{1}{L} \sum_{t=1}^{L} \mathbb{E}\big[-\log p_{h_t}(s_t) \big] \;\geq\; \log 60.$$

3.3 SHARPNESS AND THE S_5/A_5 CONTRAST

Corollary 1 (Sharpness, and the A_5 contrast)**.** *The bound is sharp as an infimum. With $s_0 = 1$, $T(x) = \operatorname{sgn} \pi(x) \in \{\pm 1\} \subset \mathrm{Spin}(3)$ and, for $0 < \eta < 1$, a readout that is continuous and positive on the orbit and places mass $1 - \eta$ uniformly on the 60 elements of the parity class recorded by the state, the loss is $\log 60 - \log(1 - \eta)$ on every word, so the infimum over continuous positive readouts is $\log 60$. With A_5 in place of S_5, $s_0 = 1$ and transitions in the binary icosahedral subgroup of $\mathrm{Spin}(3)$ lifting each token's element give a finite reachable set on which a readout, continuous and*

positive on the whole orbit, assigns $1-\eta$ *to the correct element; its loss is at most* $-\log(1-\eta)$ *at every length, for every* $0<\eta<1$.

For every fixed input law satisfying Theorem 1, the infimum over the stated pure rotor recurrences, initial states and readouts is therefore exactly $\log 60$ on S_5; this is an infimum over models, not a finite-length bound or a claim that training finds the optimum.

3.4 Fixed suffixes

Corollary 2 (Fixed-suffix readouts)**.** *Under the input and dynamics hypotheses of Theorem 1, fix an integer* $k \geq 1$ *and a decoder* q. *For every* $u \in \Sigma^k$, *let* $q(\cdot, u)$ *be continuous on* $\{gs_0 : g \in G\}$, *taking values in strictly positive probability vectors over* S_5. *With* k *and* q *fixed as* L *grows and* $z_t = (x_{t-k+1}, \ldots, x_t)$,

$$\lim_{L\to\infty} \frac{1}{L} \sum_{t=k}^{L} \mathbb{E}\big[-\log q_{h_t}(s_t, z_t)\big] \;\geq\; \log 60.$$

Corollary 2 admits the current token but not a growing context or an added recurrent memory; the affine experiments in §4.3, with writes and decay restored, lie outside the bound.

3.5 Relation to exact tracking

Exact tracking of a non-abelian target forces an algebraic factorization of the transition monoid that commuting transitions cannot satisfy (Theorem 2). Turning that algebraic map into a continuous one, the form a neural readout can use, is established when differently labelled states stay a fixed distance apart (Theorem 3), and under that hypothesis no finite number of $\mathrm{Cl}(3,0)$ or $\mathrm{Cl}(4,0)$ rotor heads tracks S_5 exactly (Proposition 3). Statements, proofs, the dense-free-subgroup counterexample and the head-count lemma are in Appendix A; its reflection-length analysis concerns represented transitions, not a learning threshold. Theorem 1 and Proposition 1 need neither exact tracking nor separation: they bound continuous predictability when exact tracking is abandoned.

4 Empirical results

The theorem isolates the predictive capacity of transition composition alone. We next study what training discovers when writes and decay are available (Table 2). The A_5 experiments establish successful task learning; the S_5 diagnostics ask whether these additional mechanisms lead to full-product prediction under the tested protocol.

Table 2: Scope of claims: the theorem establishes a predictive limit for pure transition states, whereas the empirical models utilize affine operations outside the theorem's hypotheses.

Regime	Writes (U_t)	Decay (λ_t)	Theorem applies?	Primary finding
Pure rotor	Zero	One	Yes (under stated hypotheses)	Sharp S_5 loss floor: $\log 60$
Affine models	Learned	Learned	No	A_5 learned and extrapolated; partial S_5 parity predictions

4.1 A single compact rotor layer learns A_5 and extrapolates fourfold

A one-layer $\mathrm{Cl}(3,0)$ rotor model reaches measured accuracy 1.000 on the A_5 running-product task at lengths 128, 256 and 512 after training at length 128. All three 192-head runs and both 136-head runs reach this value and confirm it at the next held-out evaluation (Table 3). The smaller configuration stores $1{,}088$ persistent scalars per layer. These results show measured generalization to four times the training length.

Table 3: A_5, one layer, local fixed-length protocol (Appendix B.1); parameters exclude embeddings; state is stored scalars per layer. First hit is the first logged accuracy 1.000 at all three lengths.

model	transition	params	state	seeds solved	first hit	acc@128/256/512
Rotor, $\mathrm{Cl}(3,0)$, 192 heads	$\mathrm{Spin}(3)$	2.36M	1,536	3/3	2.5–3k	1.00/1.00/1.00
Rotor, $\mathrm{Cl}(3,0)$, 136 heads	$\mathrm{Spin}(3)$	1.93M	1,088	2/2	3k	1.00/1.00/1.00

4.2 Solve-time variation across implemented parameterizations

We find that three parameterizations of the same ideal $\mathrm{Spin}(3)$ family—normalized quaternion, exponential map, Euler product—reach sustained success on a two-layer A_5 curriculum on most of ten seeds each, but their median solve times vary by up to roughly fivefold (Table 4, Figure 1; diagnostics in Appendix B.4). Under one training recipe, implementations of the same ideal transition family show different observed solve-time distributions. This comparison measures the combined effect of parameterization, projection and initialization choices.

Table 4: Steps to *sustained* accuracy ≥ 0.99 on A_5 (length 16): threshold held to the end of the run, ten seeds per parameterization, one protocol, every run to the 200,000-step cap. Medians and ranges condition on success; the remaining runs have no observed sustained solve.

parameterization	sustained	median steps (successes)	range
normalized quaternion	10/10	**14,500**	10k–93k
exponential map $\exp(B)$	9/10	22,000	14k–164k
Euler product $\prod_p \exp(\theta_p B_p)$	8/10	73,500	20k–186k

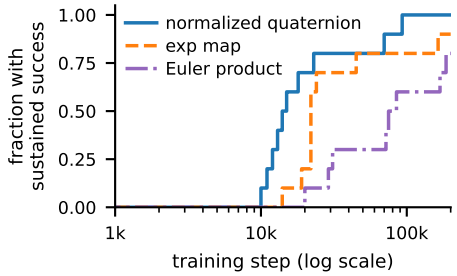


Figure 1: Same transition group and training recipe, different implemented parameterizations of $\mathrm{Spin}(3)$. A_5 with a length curriculum to 16, two layers, ten seeds per parameterization. The plot shows the fraction of all ten runs whose sustained-success onset occurs by each training step. All three parameterizations solve on most seeds; what separates them is the distribution of their solve times.

4.3 Partial solutions on S_4 and parity plateaus on S_5

We next examine whether optimization discovers available constructions for S_4 and S_5. *Availability.* S_4 is realizable through its binary-octahedral lift in $\mathrm{Spin}(3)$ with quotient decoding, and S_5 constructions become available once the family is expanded to $\mathrm{Pin}(4)$ or $\mathrm{Spin}(5)$. *Outcome.* S_4 stalls at held-out accuracy 0.316 while A_5 solves at every tested length and A_4 solves the training length but extrapolates imperfectly, without meeting the all-lengths criterion. On S_5, the tested configurations (rotors, versors, expanded groups) fail to solve the task within 30,000 steps, and the probed plateau-reaching runs settle on a parity-dominated plateau (Table 5). These runs establish non-discovery within the reported budget, not a relative sample-complexity comparison with the separate 200k curriculum study.

Table 5: S_5, one layer, identical protocol, 30,000 steps; held-out accuracy over prefixes of 128 validation sequences at lengths 128, 256 and 512 after training at 128. The sign-probed plateau-reaching runs ($\mathrm{Cl}(3,0)$, $\mathrm{Cl}(4,0)$ and $\mathrm{Cl}(5,0)$ rotor rows; the $\mathrm{Cl}(4,0)$ probe uses a reproduction checkpoint) classify the sign of the product at the training length while group-element accuracy stays low; the learned-parity versor row reaches the same loss plateau but was not sign-probed, and the fixed-parity control reaches neither. One seed per row; the learned-parity versor row is the one run in five that reached the plateau, not a seed average. These are finite-length affine models with writes and decay, outside Theorem 1, so losses below $\log 60 \approx 4.094$ do not bear on the bound.

transition	induced action	acc@128	acc@256	acc@512	loss
$\mathrm{Cl}(3,0)$ rotors	$\mathrm{SO}(3)$	0.043	0.031	0.02	3.98
$\mathrm{Cl}(4,0)$ rotors	$\mathrm{SO}(4)$	0.033	0.018	0.02	4.03
$\mathrm{Cl}(4,0)$ versors, learned parity	$\mathrm{O}(4) \supseteq S_5$	0.032	0.019	0.015	4.03
$\mathrm{Cl}(4,0)$ versors, fixed parity (no gate)	$\mathrm{O}(4) \supseteq S_5$	0.024	0.016	0.012	4.71
$\mathrm{Cl}(5,0)$ rotors, 48 heads	$\mathrm{SO}(5) \supseteq S_5$	0.051	0.030	0.017	3.96
reference: chance, $1/120$		0.008	0.008	0.008	4.79
reference: track the sign $S_5 \to C_2$ only		0.017	0.017	0.017	4.09

Interpretation. The probed plateau-reaching S_5 runs track parity while group-element accuracy stays low, and sign accuracy degrades at longer lengths; the fixed-parity control does not show this behavior, and four of five learned-parity runs do not reach the selected plateau (Appendix B.7). For the $\mathrm{Cl}(3,0)$ and $\mathrm{Cl}(4,0)$ rows these parity-dominated predictions are consistent with the parity-only floor that Theorem 1 gives for pure $\mathrm{Spin}(3)/\mathrm{Spin}(4)$ state, but they do not identify the theorem's hidden fibre in the trained state: the trained models have writes and decay, so the plateau is a parallel observation rather than a realization of the asymptotic floor. Expanded families that permit full S_5 tracking show similar partial predictions, so the shared plateau does not identify the $\mathrm{Spin}(3)/\mathrm{Spin}(4)$ obstruction as its cause. Similar sign-stage plateaus are reported for Transformers on S_5 by Zhang & Li (2026). The scopes are summarized in Table 6.

Table 6: Scope matrix. "Exact tracking" records the robust exact-tracking result (Theorem 3, Proposition 3): the S_5 exclusion requires uniform separation at all lengths in the token-local, state-readout-only model; "permits" denotes a construction with a suitable unrestricted decoder. The loss floor of Theorem 1 is separate and needs no separation. No affine failure is predicted.

task	transition family	exact tracking	affine result	interpretation
A_5	$\mathrm{Spin}(3)$	permits	solves, $4\times$ extrapolation	available and learned
S_4	$\mathrm{Spin}(3)$	permits	stalls; quotient-like loss	permitted, not learned
S_5	$\mathrm{Spin}(3), \mathrm{Spin}(4)$	excludes if separated	partial sign accuracy	empirical only
S_5	$\mathrm{Pin}(4)$	permits	partial sign accuracy	permitted, not learned
S_5	$\mathrm{Spin}(5)$	permits	partial sign accuracy	permitted, not learned
A_5	$\mathrm{Spin}(2)$ (abelian)	excludes	stalls at 3.86; outside theorem	write changes the problem

4.4 Ablation, routed replication and language-modelling scope

Removing the additive write retains the A_5 success and the S_5 failure (Appendix B.6). A routed replication in a separate codebase learns keyed A_5 tracking under a different input encoding and state-routing architecture, with five $\mathrm{SO}(3)$ and five quaternion runs at mean accuracy 0.997 and 0.994 at length 256, while a nearly parameter-matched fixed-axis commuting control reaches 0.053 (Table 11, Appendix B.3); it associates success with the richer transition family under this protocol without isolating non-commutativity, since the parameterizations also differ and stream identity is exposed in the input.

Compact rotor state is not a competitive standalone language model under the tested recipe. On 1B FineWeb-Edu tokens at approximately 125M non-embedding parameters, the rotor-only model reaches perplexity 43.83 against 30.22–33.85 for the tested matrix-state baselines, and widening the mixer from 32 to 170 heads moves perplexity by 0.3% (Table 14, Appendix C.1). Its recurrent payload is 256 stored scalars per layer against 131k–786k for the baselines; the comparison does not

isolate state size as the cause. At approximately 50M parameters, a six-rotor/six-DeltaNet hybrid improves paired DeltaNet controls in two seeds, and so does token-independent transport in its single seed (Appendix C.3): the experiments establish a gain for the tested hybrid configurations but do not identify token-dependent rotation as its source. A separate paired identity/constant intervention does not favor learned constant rotation (Appendix C.4); it uses different normalization and evaluation and is not pooled with the hybrid study. Layer allocation and width also change across hybrid conditions, and a generic-layer control is absent.

Language-model experiments delimit the scope of the group-tracking results: the standalone rotor model underperforms the tested baselines, and hybrid improvements are also obtained with token-independent rotations. These experiments establish neither competitive language modelling nor an isolated benefit from token-dependent transport.

5 Related work

Table 7: Comparison with related state-tracking analyses. Rather than establishing exact representability, our work identifies a quantitative lower bound on continuous decoding from compact state.

work	scope / assumptions	relevant result
Siems et al. (2025)	products of generalized Householder updates; published fixed-length protocol	A_5 and S_5 solved in one layer
Siems et al. (2026)	non-expansive heads, at most one non-real unit-circle eigenvalue pair per transition counted with algebraic multiplicity, finite reachability	one-layer S_5 obstruction; four-dimensional A_5 construction, not learned from random initialization
Zhang & Li (2026)	order-blind predictors under uniform i.i.d. full-group inputs	asymptotic accuracy ceiling $1/\lvert[G,G]\rvert$, implying a log-loss bound
This paper	pure finite-head $\mathrm{Spin}(3)$/$\mathrm{Spin}(4)$ state (floor) or any compact transition group (fibre formula); i.i.d. generating inputs; fixed continuous positive decoder	sharp asymptotic S_5 loss floor $\log 60$; general full-transition fibre formula $\log \lvert J\rvert$

State tracking and expressivity in linear RNNs. Restricted transitions limit recurrent computation (Merrill et al., 2024; Shakerinava et al., 2026), and Merrill et al. (2026) place permutation-diagonal and diagonal-plus-low-rank linear RNNs at NC^1 and PNC^1 under their rational-arithmetic and precision conventions. Recent architectures enlarge the transition family to reach non-abelian word problems:

- *Householder updates and their products.* Negative eigenvalues extend the groups a single generalized Householder update can reach (Grazzi et al., 2024; Yang et al., 2024); products of such updates solve A_5 and S_5 in one layer under the fixed-length protocol we adapt (Siems et al., 2025).
- *Structured and complex-diagonal transitions.* PD-SSM pairs a one-hot matrix with a complex diagonal and emulates any N-state automaton in one layer of complex dimension N (Terzic et al., 2025); Complex KDA builds one-layer trackers for the finite subgroups of $\mathrm{SO}(3)$, including a quaternionic four-dimensional A_5 construction (Siems et al., 2026).
- *Other parallel-in-time transitions.* RWKV-7 admits a one-layer construction for tracking swaps on five elements in its extended $c = 2$ setting (Peng et al., 2025); structured linear CDEs solve A_5 in one layer with longer extrapolation than ours under a variable-length protocol (Walker et al., 2025); Selective RoPE and Mamba-3 use input-dependent complex phases (Movahedi et al., 2025; Lahoti et al., 2026), sharing the commuting complex-phase transport of our $n = 2$ case (§2).

Point of departure. Concurrently, Complex KDA proves a one-layer S_5 obstruction under non-expansive head transitions, at most one non-real unit-circle conjugate eigenvalue pair per transition counted with algebraic multiplicity, and finite reachability, and reports that A_5 is representable but

not learned from random initialization under its standard setup, while construction-informed initialization succeeds under a different configuration (Siems et al., 2026); its obstruction concerns full-group inputs and permits affine updates under finite reachability. Theorem 1 answers what remains when exact tracking fails with a sharp $\log 60$ floor for any finite number of pure $\mathrm{Spin}(3)/\mathrm{Spin}(4)$ heads, and Proposition 1 gives the general full-transition infimum $\log |J|$. Their stated spectral theorem does not directly cover these rotor left actions, since a non-real unit quaternion's left action on $\mathbb{R}^4$ has each eigenvalue with multiplicity two; our result addresses asymptotic predictive loss without finite reachability, assuming i.i.d. generating inputs and fixed continuous positive decoding (Table 7).

Lie-group, rotor and Clifford representations. Lie-group parameterizations of orthogonal recurrent weights precede this work (Lezcano-Casado & Martínez-Rubio, 2019), as do rotation-representation effects on learning (Zhou et al., 2019); path development maps sequences into Lie groups (Lou et al., 2024), and Clifford networks address geometric data (Brandstetter et al., 2022; Ruhe et al., 2023; Brehmer et al., 2023; Ji, 2026), and Versor (Truong Minh Huy & Hirst, 2026) accumulates rotors in conformal geometric algebra without state-tracking tasks. We instead use rotor multiplication as the transition of an additive recurrent memory.

Group-word and finite-group tracking. Group words follow Merrill et al. (2024) and Liu et al. (2022); word problems for fixed finite non-solvable groups are NC^1-complete (Barrington, 1989); automaton decomposition is classical (Krohn & Rhodes, 1965). Measure-once quantum finite automata recognize exactly the group languages with bounded error, yet without that restriction they solve free-group word problems (Brodsky & Pippenger, 2002, Thm. 3.3, Lem. 3.5); Theorem 1 is a quantitative analogue of that robustness distinction. Zhang & Li (2026) characterize quotient solutions and derive asymptotic accuracy limits for order-blind predictors (Table 7); our theorem does not assume order-blindness. DeltaProduct explains its S_4/A_5 thresholds through $\mathrm{SO}(3)$; Howe (2026) gives the reflection budget used in Appendix A.4. The continuous-quotient step is classical topological language recognition (Steinberg, 2013). Memory-augmented and fast-weight models (Graves et al., 2014; 2016; Schmidhuber, 1992; Schlag et al., 2021) hold associations in a rapidly updated matrix, as DeltaNet's write does.

6 Limitations

Theory. The loss floor concerns token-local pure compact transitions, a fixed finite head count, i.i.d. generating inputs and a fixed continuous strictly positive decoder, with at most a fixed input suffix, and is asymptotic with no finite-length rate. The separate robust exact-tracking result instead uses exact state-only decoding and uniform separation. Neither result establishes an impossibility for affine trained models.

Experiments. The group tasks measure sampled accuracy, not general reasoning or an identified mechanism. The routed control changes the transition family and parameterization; the solve-time study conditions on success and includes projection and initialization differences. Negative S_4/S_5 outcomes are configuration- and budget-dependent, with one seed per reported S_5 table row and the learned-parity row selected from five runs. Most language comparisons have one seed; the DeltaNet hybrid has two paired seeds and incomplete causal controls; stored-state counts are analytic payloads, not measured bandwidth.

7 Conclusion

Pure finite-head $\mathrm{Spin}(3)$/$\mathrm{Spin}(4)$ state has a sharp prediction limit on S_5: under i.i.d. generating inputs, fixed continuous strictly positive readouts incur asymptotic prefix-averaged loss at least $\log 60$; parity tracking approaches the floor, whereas A_5 permits arbitrarily small loss. The general fibre principle separates the information in a transition product from what an observed state exposes.

In the related affine models, A_5 learning generalizes to four times the training length; implemented parameterizations differ in solve time, and the tested alternative constructions are not reliably discovered. The parity-dominated partial solutions leave open which memory mechanisms permit reliable full-S_5 learning.

REPRODUCIBILITY STATEMENT

Protocols, optimiser settings, budgets, seeds and evaluation samples are given in Appendices B and C. Stored configurations and logs record the experiments; crashed-and-rerun baselines and unsuccessful seeds are distinguished from completed successes. The routed replication's per-seed ledger and a verbatim snapshot of its training code are included under `results/routed_replication` and `companion_routed`. The identity/constant pair includes a frozen manifest, update journals, per-window held-out losses and a read-only evidence auditor. The 125M comparison evaluates every checkpoint on the same fixed 2.05M-token held-out sample. The loss-floor theorem, sharpness construction and fixed-suffix extension are proved in Appendix A.2; Appendix A also proves the factorization, conditional quotient theorem and subgroup lemma, with counterexamples delimiting their scope. Finite computational checks are not formal proof verification. Code, configurations, logs and the evaluation script are provided as anonymised supplementary material.

AI USE STATEMENT

Large language model assistants were used in the following ways, and the authors reviewed all AI-assisted content and take full responsibility for the submitted claims, proofs, code and text. *Writing:* to draft sections of the manuscript and to edit and polish the text. *Retrieval and discovery:* to search for and summarize related work. *Research ideation and execution:* to propose and refine hypotheses and experimental designs, to write and run experiment code and analysis scripts, and to audit the manuscript for internal consistency. *Synthetic datasets:* the group-word and control task data are generated procedurally by code written with assistant help. *Mathematical claims:* assistants helped formulate, prove and check the theorems, lemmas and propositions; the authors verified the proofs.

REFERENCES

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.

Guy E Blelloch. Prefix sums and their applications. *Technical Report CMU-CS-90-190*, 1990.

Johannes Brandstetter, Rianne van den Berg, Max Welling, and Jayesh K Gupta. Clifford neural layers for pde modeling. *arXiv preprint arXiv:2209.04934*, 2022.

Johann Brehmer, Pim de Haan, Sönke Behrends, and Taco Cohen. Geometric algebra transformer. *arXiv preprint arXiv:2305.18415*, 2023.

Emmanuel Breuillard and Tsachik Gelander. On dense free subgroups of Lie groups, 2002. URL https://arxiv.org/abs/math/0206236.

Alex Brodsky and Nicholas Pippenger. Characterizations of 1-way quantum finite automata. *SIAM Journal on Computing*, 31(5):1456–1478, 2002. arXiv:quant-ph/9903014.

A. René Geist, Jonas Frey, Mikel Zhobro, Anna Levina, and Georg Martius. Learning with 3D rotations, a hitchhiker's guide to SO(3). In *Proceedings of the International Conference on Machine Learning*, 2024. arXiv:2404.11735.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural Turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016.

Riccardo Grazzi, Julien Siems, Arber Zela, Jörg K.H. Franke, Frank Hutter, and Massimiliano Pontil. Unlocking state-tracking in linear rnns through negative eigenvalues. *arXiv preprint arXiv:2411.12537*, 2024. ICLR 2025.

Ulf Grenander. *Probabilities on Algebraic Structures*. Wiley, 1963.

Gunner Levi Howe. The automaton underneath: The additive input pathway is a parasitic attractor for state tracking in householder linear rnn. *arXiv preprint arXiv:2609.18966*, 2026. v1, submitted 16 July 2026 according to the arXiv record.

Zhongping Ji. Cliffordnet: All you need is geometric algebra. *arXiv preprint arXiv:2601.06793*, 2026.

Yukiyosi Kawada and Kiyosi Itô. On the probability distribution on a compact group. I. *Proceedings of the Physico-Mathematical Society of Japan, 3rd Series*, 22(12):977–998, 1940.

Kenneth Krohn and John Rhodes. Algebraic theory of machines. I. Prime decomposition theorem for finite semigroups and machines. *Transactions of the American Mathematical Society*, 116: 450–464, 1965.

Aakash Lahoti, Kevin Y. Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. *arXiv preprint arXiv:2603.15569*, 2026. ICLR 2026.

Mario Lezcano-Casado. Trivializations for gradient-based optimization on manifolds. In *Advances in Neural Information Processing Systems*, 2019. arXiv:1909.09501.

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 International Conference on Machine Learning*, 2019. arXiv:1901.08428.

Bingbin Liu, Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. *arXiv preprint arXiv:2210.10749*, 2022. ICLR 2023.

Hang Lou, Siran Li, and Hao Ni. Path development network with finite-dimensional lie group representation. *Transactions on Machine Learning Research*, 2024. arXiv:2204.00740.

Pertti Lounesto. *Clifford Algebras and Spinors*. Cambridge University Press, 2001.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. *arXiv preprint arXiv:2404.08819*, 2024.

William Merrill, Hongjian Jiang, Yanhong Li, Anthony Lin, and Ashish Sabharwal. Why are linear RNNs more parallelizable? *arXiv preprint arXiv:2603.03612*, 2026. ICML 2026.

Sajad Movahedi, Timur Carstensen, Arshia Afzal, Frank Hutter, Antonio Orvieto, and Volkan Cevher. Selective rotary position embedding. *arXiv preprint arXiv:2511.17388*, 2025. ICLR 2026.

Bo Peng, Ruichong Zhang, Daniel Goldstein, Eric Alcaide, Xingjian Du, Haowen Hou, Jiaju Lin, Jiaxing Liu, Janna Lu, William Merrill, et al. RWKV-7 "Goose" with expressive dynamic state evolution. *arXiv preprint arXiv:2503.14456*, 2025.

David Ruhe, Johannes Brandstetter, and Patrick Forré. Clifford group equivariant neural networks. *arXiv preprint arXiv:2305.11141*, 2023.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *ICML*, 2021.

Jürgen Schmidhuber. Learning to control fast-weight memories: an alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992.

Mehran Shakerinava et al. The expressive limits of diagonal ssms for state-tracking. *arXiv preprint arXiv:2603.01959*, 2026. ICLR 2026.

Julien Siems, Timur Carstensen, Arber Zela, Frank Hutter, Massimiliano Pontil, and Riccardo Grazzi. Deltaproduct: Improving state-tracking in linear rnns via householder products. *arXiv preprint arXiv:2502.10297*, 2025. NeurIPS 2025.

Julien Siems, Riccardo Grazzi, Korbinian Pöppel, Jaisidh Singh, Arber Zela, Timur Carstensen, Jenia Jitsev, Frank Hutter, Volkan Cevher, Antonio Orvieto, and Aaron Klein. Complex KDA: Understanding and enhancing the expressivity of Kimi Delta Attention. *arXiv preprint arXiv:2609.24797*, 2026.

Benjamin Steinberg. Topological dynamics and recognition of languages. *arXiv preprint arXiv:1306.1468*, 2013.

Karl Stromberg. Probabilities on a compact group. *Transactions of the American Mathematical Society*, 94(2):295–309, 1960.

Aleksandar Terzic, Nicolas Menet, Michael Hersche, Thomas Hofmann, and Abbas Rahimi. Structured sparse transition matrices to enable state tracking in state-space models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2509.22284.

Truong Minh Huy and Edward Hirst. Versor: A geometric sequence architecture. *arXiv preprint arXiv:2602.10195*, 2026.

Benjamin Walker, Lingyi Yang, Nicola Muca Cirone, Cristopher Salvi, and Terry Lyons. Structured linear cdes: Maximally expressive and parallel-in-time sequence models. In *Advances in Neural Information Processing Systems*, 2025. NeurIPS 2025 Spotlight; arXiv:2505.17761.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *arXiv preprint arXiv:2406.06484*, 2024.

Zhiyu Zhang and Yupeng Li. Tracking states or tracking cosets? An algebraic account of learned state tracking. *arXiv preprint arXiv:2609.29951*, 2026.

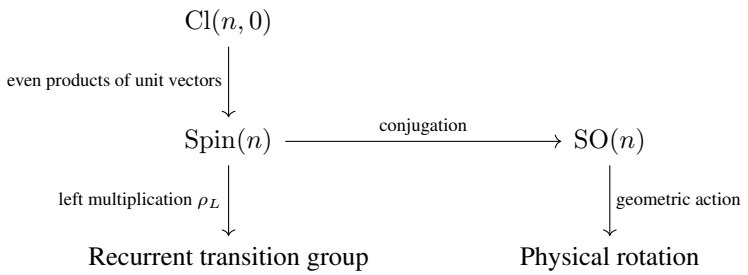
Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. arXiv:1812.07035.

Guide to the supporting evidence. The appendices follow the paper's claims. Proofs come first; experimental protocols are placed beside the results they support. The two language interventions remain separate because their initialization, normalization and evaluation differ.

claim or verification question	where to look
Sharp S_5 prediction floor and A_5 construction	Appendices A.1–A.2
Exact tracking: factorization, robustness, counterexamples, reflection length	Appendices A.3, A.4
Compact A_5 learning and tested learnability gaps	Appendix B
Curriculum sensitivity; parameterization, conditioning and head diagnostics	Appendices B.2, B.4, B.4.1, B.5
Language-memory scope and rotation controls	Appendix C
Standalone, WikiText, hybrid and identity studies	Appendices C.1, C.2, C.3, C.4
Algebra conventions, write propagation, stored state and access counts	Appendices D, D.1, D.2, D.3

A Predictive limits and supporting proofs

Transition group versus geometric action. The proofs use the left action of $\mathrm{Spin}(n)$ on the state, not the induced $\mathrm{SO}(n)$ rotation; the diagram records the distinction.



The recurrent state follows the left path and distinguishes R from $-R$; the right path would suggest, incorrectly, that the transition group is bounded by the finite subgroups of $\mathrm{SO}(n)$. The theorem concerns the left path only.

The prediction floor follows in three steps: the rotor subgroup lemma forces a nontrivial target fibre; the joint random walk converges in Cesàro average to Haar measure; the fibre then bounds the risk of a fixed continuous decoder. We prove the subgroup lemma first, followed by the loss bound and its sharpness. The exact-tracking results and their counterexamples follow as separate consequences of the same structural restriction. The final subsection treats reflection budgets, a supporting comparison rather than the new prediction theorem.

A.1 The rotor subgroup restriction

The obstruction must hold for the product of all heads, not just one head. The next lemma supplies this step for both $\mathrm{Spin}(3) \cong \mathrm{SU}(2)$ and $\mathrm{Spin}(4) \cong \mathrm{SU}(2)^2$.

The idea: no two nontrivial normal subgroups of S_5 commute elementwise, whereas the kernels of the two coordinate projections of $\mathrm{SU}(2)^m$ do, so a homomorphism onto S_5 must kill one of them and factor through fewer coordinates.

Lemma 1 (No continuous S_5 quotient of a closed subgroup of $\mathrm{SU}(2)^m$)**.** *For every $m \geq 1$ there is no closed subgroup $G \leq \mathrm{SU}(2)^m$ admitting a continuous surjective homomorphism to the discrete group S_5. Equivalently, no closed normal $N \trianglelefteq G$ gives $G/N \cong S_5$.*

Proof. We proceed by induction on m.

1. **Base case ($m = 1$):** The closed subgroups of $\mathrm{SU}(2)$ are the finite ones—cyclic C_n, binary dihedral $2D_n$, and $2T$, $2O$, $2I$ of orders 24, 48, 120—together with the maximal torus $\mathrm{U}(1)$, its normaliser, and $\mathrm{SU}(2)$ itself.

   - All the finite ones except $2I$ are solvable and S_5 is not.
   - $2I$ is perfect with normal subgroups $1, \{\pm 1\}, 2I$, so its only quotients are $2I$, A_5 and 1, while S_5 is not perfect.
   - A continuous map from a connected group to a finite discrete group is constant. Thus $\mathrm{U}(1)$ and $\mathrm{SU}(2)$ have no nontrivial continuous finite quotient, and the normaliser of the torus has only quotients of its component group C_2.

   None surjects onto S_5.

2. **Inductive step:** Let $G \leq \mathrm{SU}(2)^m$ be closed with continuous $\varphi : G \twoheadrightarrow S_5$. Write $p_{<m}$ and p_m for the projections onto the first $m-1$ factors and onto the last, and set

$$K = \ker\big(p_m|_G\big), \qquad L = \ker\big(p_{<m}|_G\big).$$

   Both are normal in G. They commute elementwise, $[K, L] = 1$, because an element of K is trivial in the last coordinate and an element of L is trivial in all the others, so their commutator is trivial in every coordinate.

   The images $\varphi(K)$ and $\varphi(L)$ are normal in S_5, whose only normal subgroups are 1, A_5 and S_5. If both were nontrivial they would both contain A_5 and, by $[K, L] = 1$, would commute elementwise; but $C_{S_5}(A_5) = 1$, so A_5 would be trivial—a contradiction. Hence one of them is trivial.

If $\varphi(L) = 1$ then φ factors through $G/L \cong p_{<m}(G) \leq \mathrm{SU}(2)^{m-1}$, and the inductive hypothesis applies. If $\varphi(K) = 1$ then φ factors through $G/K \cong p_m(G) \leq \mathrm{SU}(2)$, which is the base case. In both cases the image of the compact G is compact and hence closed. The projection is a quotient map (compact domain, Hausdorff image), so the descended homomorphism is continuous. Either way we obtain a contradiction.

□

A.2 Exact fibre loss and the rotor prediction bounds

The following proposition isolates the general principle behind the rotor bound. It optimizes the decoder for fixed transitions and a fixed input law, observing the accumulated transition itself. The distinction from observing its action on a state is explicit.

Proposition 1 (Exact fibre loss for full-transition observation)**.** *Let G be a compact metrizable group, H a finite discrete group, and x_t i.i.d. on a finite alphabet Σ, with positive probability for every symbol. Fix maps $T : \Sigma \to G$ and $\pi : \Sigma \to H$ with $\pi(\Sigma)$ generating H, and put $g_t = T(x_t) \cdots T(x_1)$, $h_t = \pi(x_t) \cdots \pi(x_1)$. Let*

$$K = \overline{\langle (T(x), \pi(x)) : x \in \Sigma \rangle} \leq G \times H, \qquad P = p_G(K), \qquad J = \{h \in H : (1, h) \in K\}.$$

Then $J \trianglelefteq H$, and $\psi(g) = hJ$ for $(g, h) \in K$ defines a continuous surjective homomorphism $\psi : P \to H/J$. With the input law and transitions fixed,

$$\inf_{q \in C(P, \Delta^\circ(H))} \lim_{L \to \infty} \frac{1}{L} \sum_{t=1}^{L} \mathbb{E}\big[-\log q_{h_t}(g_t) \big] = \log |J|,$$

where $\Delta^\circ(H)$ is the strictly positive probability simplex and each decoder q is fixed as L grows; its limit exists. For any continuous observation $F : P \to S$ into a Hausdorff space, continuous positive decoders on $F(P)$ retain this lower bound. Their infimum also equals $\log |J|$ whenever ψ factors continuously through F.

Proof. We break the proof into three components: establishing the continuous quotient, bounding the loss via Haar measure, and constructing an explicit decoder that approaches the bound.

1. **The quotient:** The projection $p_H : K \to H$ is surjective, and $J = p_H(\ker(p_G|_K))$ is normal in H. Two target elements paired with the same g differ by an element of J; conversely, multiplying by $(1, j)$ supplies every element of that coset. Thus $(g, h) \mapsto hJ$ is constant on the fibres of p_G. Since $p_G : K \to P$ is a quotient map from a compact space onto a Hausdorff space, it induces the stated continuous surjective homomorphism ψ.

2. **The lower bound:** Let ν be the law of $(T(x), \pi(x))$. Its support generates K, so $L^{-1} \sum_{t=1}^{L} \nu^{*t}$ converges weakly to normalized Haar measure m_K (Kawada & Itô, 1940, Thm. 7). For each fixed continuous positive q, the loss $\ell(g, h) = -\log q_h(g)$ is bounded and continuous on K. Its prefix-averaged expected loss therefore converges to $I_q = \int \ell \, dm_K$. Haar invariance under $(1, j)$ and Jensen's inequality give

$$I_q = \int_K \frac{1}{|J|} \sum_{j \in J} -\log q_{jh}(g) \, dm_K(g, h) \;\geq\; \log |J|,$$

   because the $|J|$ distinct labels jh have total assigned probability at most one.

3. **An explicit decoder approaching the bound:** For $0 < \varepsilon < 1$ set

$$q_h^\varepsilon(g) = (1 - \varepsilon) \frac{\mathbf{1}\{hJ = \psi(g)\}}{|J|} + \frac{\varepsilon}{|H|}.$$

   The coset indicators are continuous because H/J is finite and discrete. Hence q^ε is continuous, strictly positive and normalized. On every word, $(g_t, h_t) \in K$, so its loss is $-\log((1 - \varepsilon)/|J| + \varepsilon/|H|)$, tending to $\log |J|$ as $\varepsilon \downarrow 0$. This proves equality, with the decoder infimum outside the limit. For an observed state, composing a continuous positive decoder with F proves the lower bound. If $\psi = \bar{\psi} \circ F$ for continuous $\bar{\psi} : F(P) \to H/J$, replacing $\psi(g)$ by $\bar{\psi}(s)$ in the construction proves the same infimum.

□

Example: The rotor floor is not the abelian floor in general. On S_5 the two coincide: rotor transitions force $A_5 \subseteq J$, and $A_5 = [S_5, S_5]$ is also what abelian transitions force. However, they separate on $H = S_5 \times A_5$ with $G = \rho_L(\mathrm{Spin}(3))^{n_\mathrm{h}}$ and inputs whose labels generate H:

- **Rotor bounds:** Projecting K to $G \times S_5$ and applying the argument of Theorem 1 shows that the image of J in S_5 contains A_5. Let $N_1 \times 1 = J \cap (S_5 \times 1)$ and $1 \times N_2 = J \cap (1 \times A_5)$; both are normal in H, and by Goursat's lemma J projects onto $J_1 \supseteq A_5$ and $J_2 \leq A_5$ with $J_1/N_1 \cong J_2/N_2$. If N_1 were trivial, the case $N_2 = A_5$ would be impossible (it would force J_1 trivial), so $N_2 = 1$ and J would be the graph of an isomorphism $A_5 \to A_5$; conjugating by $(1, b)$ shows that such a graph is normal only if its image is central in A_5, which is impossible. Hence $N_1 \supseteq A_5$ and $|J| \geq 60$ for every T. Conversely, using at least two $\mathrm{Spin}(3)$ heads (or one $\mathrm{Spin}(4)$ head) to lift the A_5 factor through the binary icosahedral group and track the parity of the S_5 factor gives $J = A_5 \times 1$ and loss approaching $\log 60$ as in Corollary 1; a single $\mathrm{Spin}(3)$ head is insufficient since it cannot expose the required $C_2 \times A_5$ quotient.
- **Abelian bounds:** The rotor floor on $S_5 \times A_5$ is therefore $\log 60$ for at least two $\mathrm{Spin}(3)$ heads or one $\mathrm{Spin}(4)$ head (and $\log 120$ for a single $\mathrm{Spin}(3)$ head), whereas abelian transitions force $[H, H] = A_5 \times A_5 \subseteq J$ and a floor of $\log 3600$.

Remarks on the hypotheses. The Cesàro convergence used above is Theorem 7 of Kawada & Itô (1940); Grenander (1963) and Stromberg (1960) give modern statements. Strict positivity of the decoder is used only to keep the loss bounded: for a continuous decoder with values in $[0, 1]$, applying the argument to $\ell \wedge M$ and letting $M \to \infty$ gives the same lower bound for the $\liminf$.

The observation condition matters. Take $H = S_5$ and the parity transitions $T(x) = \mathrm{sgn}\,\pi(x)$, for which $J = A_5$. Observing g_t gives infimum $\log 60$, but with zero initial state under a linear action, $g_t s_0 = 0$ for every word. The target's Cesàro law is uniform on S_5, so the constant-state decoder has infimum $\log 120$. Thus the full-transition equality is not an equality for every recurrent-state observation. The construction proves an optimum over continuous decoders, not that a specified neural readout or training procedure attains it.

For the rotor specialization, the subgroup lemma forces $|J| \geq 60$ even before an observation is chosen. The proof below gives this specialization directly.

Write $m = n_\mathrm{h}$ or $2n_\mathrm{h}$ so that $G \cong \mathrm{SU}(2)^m$ as in Proposition 3.

Proof of Theorem 1. We apply the fibre loss bound specifically to the rotor representations.

1. **The joint group:** Let $K \leq G \times S_5$ be the closed subgroup generated by $\{(T(x), \pi(x)) : x \in \Sigma\}$. Since $G \times S_5$ is compact, the closure of the monoid generated by these pairs is already a group (as in the proof of Theorem 3), so K is also that closure. Let p_G and p_H be the coordinate projections; $P = p_G(K)$ is a closed subgroup of G, and $p_H(K) = S_5$ because π generates S_5. Put $J = \{j \in S_5 : (1, j) \in K\} = p_H(\ker p_G)$. Since $\ker p_G \trianglelefteq K$ and p_H maps K onto S_5, $J \trianglelefteq S_5$.

2. **J is nontrivial:** If $J = 1$ then p_G is injective on K, hence a homeomorphism from the compact K onto the Hausdorff P, and $p_H \circ p_G^{-1} : P \to S_5$ is a continuous surjective homomorphism from a closed subgroup of $\mathrm{SU}(2)^m$ onto S_5, contradicting Lemma 1. As the normal subgroups of S_5 are 1, A_5 and S_5, $|J| \geq 60$.

3. **Cesàro convergence:** Let ν be the law of $(T(x), \pi(x))$; it charges every generator, so its support generates K as a closed group. The law of (g_t, h_t) is the t-fold convolution ν^{*t} (the factors are i.i.d., so the order of multiplication does not change the law of the product). By the mean ergodic theorem of Kawada & Itô (1940, Thm. 7), $\bar{\nu}_L = \frac{1}{L}\sum_{t=1}^{L} \nu^{*t}$ converges weakly to the normalized Haar measure m_K. Define $\ell(g, h) = -\log p_h(g s_0)$ on K. Because p is continuous and strictly positive on the compact orbit, ℓ is bounded and continuous, so $\frac{1}{L}\sum_{t=1}^{L} \mathbb{E}\,\ell(g_t, h_t) = \int \ell \, d\bar{\nu}_L \to \int \ell \, dm_K =: I$; in particular the limit exists.

4. **The floor:** For $j \in J$, $(1, j) \in K$ and m_K is invariant under left translation by it, so $I = \int \ell(g, jh)\, dm_K(g, h)$ for every $j \in J$, and averaging,

$$I = \int \frac{1}{|J|} \sum_{j \in J} -\log p_{jh}(gs_0)\, dm_K(g, h).$$

For fixed (g, h) the labels jh, $j \in J$, are $|J|$ distinct elements of S_5, so $\sum_{j \in J} p_{jh}(gs_0) \le 1$, and Jensen's inequality gives

$$\frac{1}{|J|} \sum_{j \in J} -\log p_{jh}(gs_0) \;\ge\; -\log\Big(\frac{1}{|J|} \sum_{j \in J} p_{jh}(gs_0)\Big) \;\ge\; \log|J| \;\ge\; \log 60.$$

Integrating against m_K gives $I \ge \log 60$.

□

Proof of Corollary 1. Both constructions use $s_0 = 1$, so the orbit of G is G itself and the reachable states form a finite subset F of it. A readout prescribed on F extends to a continuous positive readout on G: choose pairwise disjoint open neighbourhoods U_f of the points $f \in F$ and continuous $\beta_f : G \to [0, 1]$ with $\beta_f(f) = 1$ and support in U_f, and set $p(s) = \sum_f \beta_f(s)\, q_f + \big(1 - \sum_f \beta_f(s)\big)\, u$, where q_f is the prescribed positive distribution at f and u is uniform; p is continuous, positive, and equals q_f at f. For the parity tracker, $g_t = \mathrm{sgn}(h_t) \in \{\pm 1\}$ records exactly the parity of h_t; with $q_{\pm 1}$ placing $(1-\eta)/60$ on each element of the matching parity class and $\eta/60$ on each of the others, every word has loss $\log 60 - \log(1-\eta)$, whose infimum over $0 < \eta < 1$ is $\log 60$ and is not attained by a positive readout. For A_5, let $2I \subset \mathrm{Spin}(3)$ be the binary icosahedral group, which covers A_5 two-to-one, and choose $T(x) \in 2I$ with image $\pi(x)$; then $g_t \in 2I$, h_t is the image of g_t, $F = 2I$, and q_g placing $1 - \eta$ on the image of g and $\eta/59$ on the other labels gives loss $-\log(1-\eta)$ at every length. □

Proof of Corollary 2. Use the joint group K, increment law ν, Haar measure m_K and hidden subgroup J from the proof of Theorem 1. For $u = (u_1, \ldots, u_k)$ let $c(u) = (T(u_k) \cdots T(u_1), \pi(u_k) \cdots \pi(u_1)) \in K$. Conditional on $z_t = u$, the preceding $t - k$ inputs remain i.i.d. and independent of u, so (g_t, h_t) has law $\delta_{c(u)} * \nu^{*(t-k)}$. The loss $\ell_u(g, h) = -\log q_h(gs_0, u)$ is bounded and continuous on K. Cesàro convergence and left invariance of Haar measure therefore give $L^{-1} \sum_{t=k}^{L} \mathbb{E}[\ell_u(g_t, h_t) \mid z_t = u] \to \int \ell_u\, dm_K =: I_u$; including ν^{*0} and the normalization $(L - k + 1)/L \to 1$ do not change the limit. For each u, the same averaging over $(1, j)$, $j \in J$, and Jensen argument gives $I_u \ge \log|J| \ge \log 60$. Averaging over the finitely many suffixes with their time-independent probabilities proves the stated limit and bound. □

The loss bounds need neither exact tracking nor a separation margin. They concern pure transition-only dynamics, allowing the decoder a fixed input suffix but not additional recurrent memory or context growing with length. Writes and decay remain outside the hypotheses, and the bounds concern asymptotic prefix averages rather than a fixed length.

A.3 Exact tracking: factorization, robustness and counterexamples

Exact decoding forces an algebraic quotient. Extending it continuously to a compact closure requires the additional separation hypothesis; the counterexamples below show why that hypothesis cannot be omitted.

Proposition 2 (Algebraic factorization and its robust compact-group consequence)**.** *Consider the* transition-only *recurrence $S_t = T(x_t)\, S_{t-1}$, obtained from equation 2 by setting $\lambda_t = 1$, $U_t = 0$, with token-local transitions and a state-only decoder. Exact tracking of a finite group H for every word implies a surjective monoid homomorphism from the generated transition monoid M to H (Theorem 2). If the ambient linear group G is compact and there is a single $\delta > 0$ separating all pairs of reachable states with different target labels, then H is a continuous quotient of the closed subgroup $\overline{M} \le G$ (Theorem 3). For rotor heads, $G = \rho_L(\mathrm{Spin}(n))^{n_h}$. Compactness alone does not give this second conclusion.*

The ideal model: consider the transition-only recurrence obtained from equation 2 with $\lambda_t = 1$ and $U_t = 0$: a token-local map $T : \Sigma \to G$ into a linear group on V gives $S(w) = T(w_k) \cdots T(w_1) S_0$ with fixed S_0 and a state-only decoder $f : V \to H$; the target symbols $\pi : \Sigma \to H$ generate the finite group H, and $\pi(w) = \pi(w_k) \cdots \pi(w_1)$, including the empty word.

Theorem 2 (Exact tracking forces a monoid factorization)**.** *Suppose the recurrence tracks the word problem exactly: $f(S(w)) = \pi(w)$ for every word w. Let $M = \{T(w_k) \cdots T(w_1) : w \in \Sigma^*\}$, including the identity. There is a unique surjective monoid homomorphism $\varphi : M \to H$ with $\varphi(\rho(w)) = \pi(w)$, where $\rho(w) = T(w_k) \cdots T(w_1)$. Thus*

$$H \cong M/\sim_\varphi, \qquad a \sim_\varphi b \iff \varphi(a) = \varphi(b).$$

Here $\sim_\varphi$ is the kernel congruence*. The result also holds for a monoid of linear operators, without invertibility or compactness.*

Proof. If $\rho(w) = \rho(v)$ then $S(w) = S(v)$, so exact decoding gives $\pi(w) = \pi(v)$. Consequently, $\varphi(\rho(w)) := \pi(w)$ is well-defined and unique. With ordinary concatenation, $\rho(uv) = \rho(v)\rho(u)$ and $\pi(uv) = \pi(v)\pi(u)$: both word maps reverse multiplication order.

For $a = \rho(u)$ and $b = \rho(v)$ we therefore have:

$$\varphi(ab) = \varphi(\rho(vu)) = \pi(u)\pi(v) = \varphi(a)\varphi(b).$$

The empty word maps to the identity. Since H is finite, its generating set also generates it as a monoid (inverses are positive powers), giving surjectivity. The equivalence classes of $\sim_\varphi$ multiply well-definedly and map bijectively to H, proving the isomorphism. □

This is an algebraic factorization, not a closed-subgroup quotient. In particular, a commuting monoid M has only commuting homomorphic images, so this transition-only model cannot exactly track a non-abelian target group. Neither this conclusion nor the factorization requires a complexity assumption.

Remark 1 (Why compactness alone is insufficient)**.** *A monoid's identity fibre alone is not the definition of its kernel congruence. To see why compactness alone does not force a continuous quotient, consider two counterexamples:*

- ***No continuous extension:*** *Let $a = e^{2\pi i\alpha} \in \mathrm{U}(1)$ with α irrational. For $M = \{a^n : n \geq 0\}$ the assignment $\varphi(a^n) = n \bmod 2$ is a well-defined surjective monoid homomorphism to C_2. Both the even and odd powers are dense in $\mathrm{U}(1)$, so it has no continuous extension to the closure. This refutes automatic extension.*
- ***Unrestricted S_5 tracking:*** *There is a dense free subgroup $F = \langle a, b\rangle \cong F_2$ in $\mathrm{SU}(2)$ (Breuillard & Gelander, 2002). Map a to a 5-cycle and b to a transposition generating S_5; freeness gives an epimorphism $F \twoheadrightarrow S_5$. Use a, b, a^{-1}, b^{-1} as unit-quaternion transitions, initial state 1, and this epimorphism as the readout on the reachable states F (extend it arbitrarily elsewhere).*

This second construction is an exact transition-only tracker allowed by the first theorem, although Lemma 1 excludes a continuous S_5 quotient of any closed subgroup of $\mathrm{SU}(2)$. The decoder is discontinuous and not uniformly robust. This proves why an arbitrary-readout impossibility claim is false, and why the uniform separation hypothesis is required.

Theorem 3 (Uniform separation gives a continuous compact-group quotient)**.** *In the setting of Theorem 2, suppose $G \leq \mathrm{GL}(V)$ is compact, V is finite-dimensional, and differently labelled reachable states are uniformly separated: there exists $\delta > 0$ such that, for all finite words u, v,*

$$\pi(u) \neq \pi(v) \quad \Longrightarrow \quad \|S(u) - S(v)\| \geq \delta.$$

Then $K = \overline{M}$ is a compact subgroup of G, and φ extends to a continuous surjective homomorphism $\bar{\varphi} : K \to H$. In particular $H \cong K/N$ with $N = \ker \bar{\varphi}$ a closed normal subgroup.

Proof of Theorem 3. We proceed in three steps:

1. **Closure as a group:** The closure K is a compact submonoid. It is also a group: for $x \in K$, compactness provides powers x^{n_j} converging to the same limit along a subsequence with $n_{j+1} - n_j \geq 2$. Then $x^{n_{j+1}-n_j} \to 1$ and $x^{n_{j+1}-n_j-1} \to x^{-1}$, so closedness puts x^{-1} in K.

2. **Disjoint Fibres:** For each $h \in H$ put $K_h = \overline{\varphi^{-1}(h)} \subseteq K$. These sets cover K because H is finite. They are pairwise disjoint: if x belonged to $K_h \cap K_{h'}$, sequences $a_j, b_j \in M$ of those labels would both converge to x, giving $\|a_j S_0 - b_j S_0\| \to 0$, which contradicts uniform separation when $h \neq h'$.

3. **Continuous Homomorphism:** The finite closed partition (K_h) is also relatively open. Define $\bar{\varphi}(x) = h$ on K_h; this map is continuous, surjective, and extends φ. If $x \in K_h$ and $y \in K_{h'}$, approximate them by elements of the corresponding fibres in M. Their products converge to xy and have label hh', so $xy \in K_{hh'}$. Thus $\bar{\varphi}$ is a homomorphism. Its kernel is closed and normal, giving the quotient.

□

Uniform separation is an additional hypothesis, not a consequence of exact argmax accuracy. It holds, for example, when reachable states form a finite correctly decoded set, or when continuous class scores have a uniform positive classification margin on the compact orbit closure. Positive margins on a finite evaluation sample do not establish it for all words. The hypothesis is not imposed on the whole vector space, only on differently labelled reachable states. None of our experiments certifies it for unbounded sequence length.

Why the trained affine model is outside the exact-tracking obstruction. Writes can distinguish words with the same rotor product, and scalar decay changes the transition group. An affine update can still be made linear in homogeneous coordinates:

$$\begin{pmatrix} S_t \\ 1 \end{pmatrix} = \begin{pmatrix} T_t & U_t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} S_{t-1} \\ 1 \end{pmatrix}.$$

With token-local updates and state-only decoding, the algebraic factorization applies to this augmented monoid. It is not the compact left-rotor group, however, so the compact-group exclusion does not follow. These exact-tracking statements also assume a state-only decoder, unlike the fixed-suffix loss bound of Corollary 2. Therefore, the rotor exclusions do not explain the trained affine models' failures.

Corollary 3 (Uniformly separated exact tracking with $\mathrm{Cl}(3,0)$ rotors)**.** *Let a transition-only recurrence have n_h heads whose transitions are rotors of $\mathrm{Cl}(3,0)$, so that the transition representation is $\rho_L(\mathrm{Spin}(3))^{n_\mathrm{h}} \cong \mathrm{SU}(2)^{n_\mathrm{h}}$. If it tracks the word problem of a finite group H exactly and satisfies the uniform separation hypothesis of Theorem 3, then H is a continuous quotient of a closed subgroup of this representation.*

In particular, A_5 is permitted through the binary icosahedral subgroup of $\mathrm{Spin}(3)$ and its quotient, whereas S_5 is excluded by Lemma 1.

Proof of Corollary 3. Theorem 3 applies with $G = \rho_L(\mathrm{Spin}(3))^{n_\mathrm{h}} \cong \mathrm{SU}(2)^{n_\mathrm{h}}$. The binary icosahedral subgroup of $\mathrm{Spin}(3)$ surjects onto A_5, while Lemma 1 rules out a continuous S_5 quotient of a closed subgroup of any such power. The A_5 construction uses a finite reachable subgroup, initial state 1, and its quotient readout; different labels are separated because the reachable set is finite. □

A closely related extension principle appears in topological language recognition, where a clopen subset of a compact monoid has an open syntactic congruence (Steinberg, 2013, Thm. 2.6 and the proof of Prop. 2.8); our use of it is rotor-specific. For rotor heads G is a faithful copy of $\mathrm{SU}(2)^{n_\mathrm{h}}$ or $\mathrm{SU}(2)^{2n_\mathrm{h}}$, and no closed subgroup of $\mathrm{SU}(2)^m$ has a continuous S_5 quotient (Lemma 1):

Proposition 3 (Conditional S_5 obstruction at any finite head count)**.** *Let a transition-only recurrence have n_h heads whose transitions are rotors of $\mathrm{Cl}(3,0)$ or of $\mathrm{Cl}(4,0)$, so that the transition representation is $\rho_L(\mathrm{Spin}(3))^{n_\mathrm{h}}$ or $\rho_L(\mathrm{Spin}(4))^{n_\mathrm{h}}$. It cannot track S_5 exactly for all words while*

satisfying the uniform state-separation hypothesis of Theorem 3, for any finite n_{h}. Without that hypothesis an arbitrary-readout impossibility claim is false (Appendix A, Remark 1); restoring decay or the write is outside this proposition.

Proof. If such tracking were possible, Theorem 3 would produce a closed subgroup G of $\rho_L(\mathrm{Spin}(n))^{n_{\mathrm{h}}}$ ($n \in \{3, 4\}$) with a continuous surjection $\varphi : G \twoheadrightarrow S_5$. The left-multiplication representation is faithful, so we may identify G with a closed subgroup of $\mathrm{Spin}(n)^{n_{\mathrm{h}}}$. Since $\mathrm{Spin}(3) \cong \mathrm{SU}(2)$ and $\mathrm{Spin}(4) \cong \mathrm{SU}(2) \times \mathrm{SU}(2)$, this ambient group is $\mathrm{SU}(2)^m$ with $m = n_{\mathrm{h}}$ or $m = 2n_{\mathrm{h}}$. Lemma 1 applies. □

Remark 2 (The number of heads matters, and an experiment that shows it)**.** *Stating these results for a single copy of the induced geometric $\mathrm{SO}(n)$ action would be too weak, and in the case of Corollary 3 actually false. With n_{h} heads the transition group is a* product*, and groups outside the list of finite subgroups of $\mathrm{SO}(3)$ can embed in the product of induced actions subdirectly. The elementary abelian group $(\mathbb{Z}_2)^3$ is the smallest example: it is not cyclic, not dihedral (D_4 is non-abelian), and not a quotient of A_4, S_4 or A_5, yet its three coordinate projections $(\mathbb{Z}_2)^3 \to \mathbb{Z}_2 \hookrightarrow \mathrm{SO}(3)$ have trivially intersecting kernels, so its induced action embeds in $\mathrm{SO}(3)^3$. Empirically a 192-head $\mathrm{Cl}(3, 0)$ model reaches accuracy 1.000 on the $(\mathbb{Z}_2)^3$ word problem at the training length within $1{,}000$ steps; the stricter all-lengths solve criterion of Appendix B.1 is met at 26k steps. The criterion constrains the transition group of the* whole layer*, not of one head; the $(\mathbb{Z}_2)^3$ experiment shows that the single-head reading is insufficient. Lemma 1 supplies the* conditional *S_5 obstruction at any finite head count; this experiment does not establish uniform separation for all words.*

The product argument uses the commuting coordinate kernels and the normal-subgroup structure of S_5; it is not a general rule inferred from a list of simple composition factors. The same obstruction does not apply to $\mathrm{Pin}(4)$ or $\mathrm{Spin}(5)$: the respective geometric images $\mathrm{O}(4)$ and $\mathrm{SO}(5)$ contain S_5, and the finite preimage of this subgroup under the double cover surjects onto it. This gives a transition-only construction with a finite separated orbit and an unrestricted state readout, not a guarantee for a particular neural readout or optimizer.

A.4 Reflection length and represented transitions

A separate question concerns the number of reflections needed to implement one represented transformation. For $Q \in \mathrm{O}(n)$, the reflection length $\ell(Q) = \mathrm{rank}(Q - I)$ is the minimum number of hyperplane reflections whose product is Q. For rotors, this section measures the induced action on vectors, not left multiplication on the recurrent state. These differ: a non-identity $\mathrm{Cl}(3, 0)$ rotor has left-action reflection length four on the even subalgebra and eight on the allocated multivector, but induces a vector rotation of length at most two (-1 induces identity).

Lemma 2 (Rotor budget)**.** *If $\Omega \in \mathfrak{so}(n)$ has rank $2r$ then $\ell(\exp \Omega) \le 2r$, with equality generically. A bivector of $\mathrm{Cl}(n, 0)$ has rank at most $2\lfloor n/2 \rfloor$, so a rotor transition in $\mathrm{Cl}(n, 0)$ has $\ell \le b := 2\lfloor n/2 \rfloor$: $b = 2$ for $n = 3$ and $b = 4$ for $n = 4, 5$.*

Proof of Lemma 2. $\exp \Omega$ fixes $\ker \Omega$ pointwise, a subspace of dimension $n - 2r$, so $\ker(\exp \Omega - I)$ has dimension at least $n - 2r$. On its r nonzero skew-symmetric two-dimensional blocks, the exponential has no fixed vector unless a rotation angle is a multiple of 2π; excluding these exceptional angles gives equality. A product of n_h literal Householder reflections likewise has $\ell \le n_h$. This bounds a specified operator, not all products of its generators. □

Proposition 4 (Averaging bound)**.** *Let $G \subset \mathrm{O}(V)$ be finite, $V \neq 0$, with $V^G = 0$, and let the nonnegative integer b satisfy $\ell(g) \le b$ for all $g \in G$. Then $\dim V \le 2b - 1$.*

Proof of Proposition 4. Write $n = \dim V$ and $\chi(g) = \mathrm{tr}\, g$. For orthogonal g, $\mathrm{im}(g - I) \perp \ker(g - I)$, so $V = V^g \oplus W_g$ with $W_g = \mathrm{im}(g - I)$. We have $\chi(g) = (n - \ell(g)) + \mathrm{tr}(g|_{W_g}) \ge n - 2\ell(g) \ge n - 2b$, since $g|_{W_g}$ is orthogonal on an $\ell(g)$-dimensional space. Averaging over the group, $\sum_g \chi(g) = |G| \dim V^G = 0$, hence $\sum_{g \neq 1} \chi(g) = -n < 0$. If $n \ge 2b$ every term is ≥ 0, a contradiction. □

Corollary 4. *A finite group admits a faithful orthogonal representation with $\ell(g) \leq 2$ for all g iff it embeds in $\mathrm{SO}(3)$, i.e. is cyclic, dihedral, A_4, S_4 or A_5. In particular $\ell = 2$ is the floor of the non-abelian regime: $\ell \leq 1$ throughout forces G to be trivial or C_2.*

Proof of Corollary 4. The trivial group is immediate. Otherwise discarding V^G preserves faithfulness and leaves a nonzero representation. Proposition 4 gives $n \leq 3$; adding trivial coordinates embeds it in $\mathrm{O}(3)$ without changing reflection lengths. The map $\psi(g) = \det(g)\, g$ is a homomorphism $\mathrm{O}(3) \to \mathrm{SO}(3)$ with kernel $\{\pm I\}$, and $-I \notin G$ since $\ell(-I) = 3$, so $\psi|_G$ is injective. Conversely every non-identity rotation of $\mathbb{R}^3$ has $\ell = 2$. For $b = 1$ the bound gives $n \leq 1$. □

Lemma 3 (S_5 needs four)**.** *In every faithful real representation ρ of S_5, $\max_g \ell(\rho(g)) \geq 4$, attained by $\rho = \mathrm{std} \oplus \mathrm{sgn}$ in $\mathrm{SO}(5)$ (Table 8).*

Proof of Lemma 3. Let σ be a 5-cycle. Over $\mathbb{C}$, $\rho(\sigma)$ is diagonalizable with fifth roots of unity as eigenvalues. All four nontrivial powers of σ are conjugate in S_5, so raising eigenvalues to any nonzero power modulo five preserves their multiplicities. The four primitive fifth roots therefore have equal multiplicity. Faithfulness makes this multiplicity positive, so $\mathrm{rank}(\rho(\sigma) - I) \geq 4$. Conversely, $\mathrm{std} \oplus \mathrm{sgn}$ is faithful and lies in $\mathrm{SO}(5)$; every odd-dimensional proper rotation has an eigenvalue one, hence reflection length at most four. A 5-cycle attains four, establishing sharpness without a character-table assumption. □

A generator budget is not a full-group budget. Let the alphabet consist of the four adjacent transpositions generating S_5 and let T_i be their permutation matrices restricted to the standard four-dimensional representation. Each T_i is a single reflection, so $\ell(T_i) = 1$. Nevertheless their products form a faithful copy of S_5, and a 5-cycle has reflection length four. Storing the accumulated matrix, initialized to identity, and decoding its group element gives an exact, finite-state, uniformly separated tracker with one reflection per generator token.

Consequently, $\min_\rho \max_{g \in G} \ell(\rho(g))$ is not a necessary per-token bound for an arbitrary generating alphabet. It is a necessary budget when every group element must itself be available as one step in a faithful representation. Generalized delta-rule factors and quotient-based decoders require further analysis beyond that representation setting.

Table 8 lists reflection lengths by class, and Table 9 sets the resulting budgets beside DeltaProduct's published thresholds.

Table 8: Reflection length by conjugacy class of S_5 (all seven classes) in the natural, standard and standard $\oplus$ sign representations. Odd classes pay for parity in $\mathrm{SO}(5)$; the maximum of 4 is forced by the 5-cycles in every faithful representation (Lemma 3).

class	size	ℓ (nat)	ℓ (std)	sign	ℓ (std$\oplus$sgn)
$(1,1,1,1,1)$	1	0	0	$+$	0
(5)	24	4	4	$+$	4
$(4,1)$	30	3	3	$-$	4
$(3,2)$	20	3	3	$-$	4
$(3,1,1)$	20	2	2	$+$	2
$(2,2,1)$	15	2	2	$+$	2
$(2,1,1,1)$	10	1	1	$-$	2

For a faithful direct-sum orthogonal representation, reflection lengths add over heads. After removing its common fixed subspace, the averaging bound applies to that effective representation, not automatically to the allocated state dimension. Neither this calculation nor the conditional compact-group result supplies an optimizer-independent threshold for the affine architectures evaluated here.

B GROUP-TRACKING PROTOCOLS AND EVIDENCE

These experiments ask whether compact tracking is learned, and how sensitive it is to the implemented transition family and training protocol. Each subsection pairs its protocol with the corre-

Table 9: DeltaProduct's published single-layer thresholds (Siems et al., 2025, Fig. 5) against the minimum over faithful real representations of the maximal reflection length, and against the $n-1$ Householder count for permutations of n elements (Grazzi et al., 2024). Reflection length numerically matches all four; this is a descriptive comparison, not a theorem predicting the learned thresholds. The permutation construction is loose on S_4 and A_5, the two cases Siems et al. (2025) explain by their embedding in $\mathrm{SO}(3)$. The representation-relative comparison is also discussed by Howe (2026).

group	$\lvert G\rvert$	$\min_\rho \max_g \ell$	published n_h	$n-1$ count
S_3	6	2	2	2
S_4	24	2	2	3
A_5	60	2	2	4
S_5	120	4	4	4

sponding evidence. The routed replication uses a separate architecture and recipe; its seeds are not pooled with the single-layer results.

For the group experiments, unless overridden below, defaults are AdamW $(\beta_1, \beta_2) = (0.9, 0.98)$, gradient clipping at 1.0, linear warm-up and fp32. Fixed held-out batches use a separate seed. Accuracy is the fraction of prefixes whose running group product is predicted correctly; chance is $1/|G|$. Unless overridden below, training stops after two consecutive evaluations reach accuracy ≥ 0.999 at every tested length; solve time is the first such evaluation.

B.1 Local fixed-length group protocol (§4.1, §4.3)

Tokens are sampled uniformly from the whole group, and each target is the running product: this is group-element classification, not tracking the position of one object. Howe (2026) reports different Householder thresholds for point tracking than Siems et al. (2025) reports for group-element classification. Comparisons therefore need the task format as well as the group.

Our experiment adapts DeltaProduct's fixed-length protocol (Siems et al., 2025, App. C.1); it does not reproduce its published budget. The important differences are:

setting	published DeltaProduct	local comparison
training examples	2M fixed sequences, 100 epochs	sampled online
training length / batch	128 / 1,024	128 / 256
AdamW β_2	0.999	0.98
learning rate / weight decay	10^{-3} / 10^{-6}	10^{-3} / 10^{-6}
schedule	cosine	1k warm-up, cosine to each cap
training cap	200M sequences	rotor 12k steps; baseline 60k
baseline runs completed	—	three seeds, 60k steps each

Two baseline runs crashed at 40k and 38k steps and were rerun from scratch to the cap; all three completed runs are reported, and none solves: held-out accuracy at the cap is 0.039–0.041 at length 128 and 0.022–0.023 at length 512 (logs and the crashed attempts in `results/a5_dp_fullbudget`). Completing the local 60k budget is not a reproduction of the published budget: even the 60k cap supplies only about 7.7% of the published sequence budget, and the different local caps produce different learning-rate horizons. Neither a published-recipe reproduction nor a general sample-efficiency ratio follows; all three local DeltaProduct baseline runs did not reach the solve criterion within these truncated budgets.

Architecture & Evaluation Details:

- **Dimensions:** Both local models have one layer and width 384, plus 23,040 tied embedding parameters.
- **Evaluation:** 128 held-out sequences at lengths 128, 256, and 512, normally evaluated every 500 steps (every 1,000 in one rotor run). Accuracies are fractions of $128 \times L$ prefixes, logged to three decimals (e.g., 1.000 means $\geq 99.95\%$ correct).

- **Rotor Model:** The $\mathrm{Cl}(3,0)$ rotor uses either 192 heads (2.36M non-embedding parameters) or 136 heads (1.93M parameters, matching DeltaProduct's 1.93M). It features input-dependent normalized quaternions, a learned initial state, normalized q, k, and an additive write $(1-\lambda_t)k_t v_t$.
- **DeltaProduct Model:** Uses two generalized Householder factors, 12 heads of dimension 32, and $\beta \in (0,2)$, allowing negative eigenvalues.

Group-Specific Adjustments:

- **S_5:** Uses the same recipe with a 30k cap. The $\mathrm{Cl}(3,0)$ arm uses 192 heads; the $\mathrm{Cl}(4,0)$ arm uses 48 heads to reduce geometric-product memory.
- **$(\mathbb{Z}_2)^3$, S_4, and A_4:** Use 192 $\mathrm{Cl}(3,0)$ heads and the 30k cap. $(\mathbb{Z}_2)^3$ stops on the solve criterion at 26k; S_4 and A_4 reach the cap without hitting the all-lengths solve criterion (though A_4 reaches accuracy 1.000 at the training length).
- **Abelian Control:** The A_5 control uses `vec_dim 2`, 1.85M parameters, and the 30k cap.

B.2 Sensitivity to the length curriculum

This experiment tests sensitivity to the curriculum rule and seed under a second training protocol.

The model and evaluation follow Appendix B.1, with batch 64, learning rate 3×10^{-3} (flat after 1,000 warm-up steps), weight decay 10^{-2}, 200,000-step cap. The training length starts at 2 and doubles $(2, 4, \ldots, 128)$ whenever the training accuracy over the last 200 steps exceeds 0.95 ("window" rule) or, in the alternative rule, exceeds 0.95 for 200 consecutive steps ("consecutive"). The two rules have different outcomes for DeltaProduct; both are reported below.

Table 10: A_5 with a doubling length curriculum to 128, one layer. All runs to a 200k-step cap.

model	curriculum rule	seed	reached L=128	acc@128/256/512
Rotor $\mathrm{Cl}(3,0)$	window	0	yes	1.00 / 1.00 / 1.00
Rotor $\mathrm{Cl}(3,0)$	window	1	yes	1.00 / 1.00 / 1.00
Rotor $\mathrm{Cl}(3,0)$	window	2	yes	1.00 / 1.00 / 1.00
Rotor $\mathrm{Cl}(3,0)$	window	3	yes	1.00 / 1.00 / 1.00
Rotor $\mathrm{Cl}(3,0)$	consecutive	0	yes	1.00 / 1.00 / 1.00
DeltaProduct n_h=2	window	0	no (stalled at L=32)	0.38 / 0.21 / 0.11
DeltaProduct n_h=2	window	1	yes	0.999 / 0.98 / 0.86
DeltaProduct n_h=2	window	2	no (stalled at L=8)	0.07 / 0.04 / 0.02
DeltaProduct n_h=2	consecutive	0	yes	0.97 / 0.92 / 0.74

Two observations. First, the rotor model reaches 1.000 at every length on all five runs, under both advance rules. Second, the baseline is strongly sensitive to the schedule and the seed: the same model, seed and budget gives 0.97 under one advance rule and 0.38 under the other, and one seed in three reaches the training length under the window rule. This sensitivity is why the main text uses a reduced fixed-length adaptation instead. Table 10 (Appendix B.2) reports all listed runs; it does not replace their distribution with a best-run comparison.

B.3 Routed compact-state replication (§4.4)

Companion-study protocol. Keyed A_5 tracking uses four generator/inverse symbols, rather than sampling the whole group as in Appendix B.1. Each input encodes the stream key and symbol; its label is that stream's running product.

- **Architecture:** Each of four banks holds four rotation blocks, initialized at identity: 144 stored scalars for $\mathrm{SO}(3)$ or 64 for quaternions. A learned straight-through hard router selects a bank from the token embedding.
- **Updates:** The update left-multiplies each selected block by a symbol-derived rotation; a learned gate scales the angle, not an interpolation with the old state. The readout concatenates the current token embedding with the committed bank's four rotation matrices (also for quaternions, through their sign-invariant $\mathrm{SO}(3)$ action).

Table 11: Routed compact-state replication (companion study; different architecture and protocol from Table 3). Keyed A_5 tracking with four state banks, training lengths uniform on $[3, 40]$, five seeds; held-out accuracy at length 256 (chance $1/60 \approx 0.017$). The $\mathrm{SO}(3)$ and unit-quaternion rows are two parameterizations of the same rotation transitions in the companion code, both read out through their $\mathrm{SO}(3)$ action.

model	params	mean acc@256	seeds > 0.97	transition
$\mathrm{SO}(3)$ bank	12,641	0.997	5/5	non-commutative
unit-quaternion bank	12,641	0.994	5/5	non-commutative
fixed-axis control	12,609	0.053	0/5	commuting

- **Control:** The commuting control uses one fixed axis with one angle per symbol per block and stores 144 scalars.
- **Training:** Uses lengths uniform on $[3, 40]$, batch 128, AdamW with constant learning rate 2×10^{-3}, weight decay 10^{-2} and default betas $(0.9, 0.999)$, with no warm-up, scheduler or clipping. Each of seeds 0–4 runs 5,000 CPU updates, then evaluates on 256 sequences at each length 40, 128 and 256, averaging accuracy over all prefixes.
- **Parameters:** 12,641 ($\mathrm{SO}(3)$, quaternion) and 12,609 (control).

Means and counts above 0.97 in Table 11 summarize the included ledger; that cutoff is not the original script's printed success threshold. A separate CPU launcher preserves the seven archived source files and writes reproduction results to a new output path.

B.4 Parameterization comparison and diagnostics

Experimental Setup:

- **Task:** A_5 with a $+1$ length curriculum from 2 to 16 (threshold 0.95 over 50 steps).
- **Model:** Two layers, d_{model}=256, 128 heads of $\mathrm{Cl}(3, 0)$.
- **Training:** Batch 128, learning rate 3×10^{-3}, 1,000 warm-up steps, 200,000-step cap.
- **Evaluation:** Held-out evaluation every 1,000 steps at lengths 16 and 32. Steps-to-threshold is the first evaluation after which accuracy at length 16 stays ≥ 0.99 for the rest of the run. Ten seeds per parameterization under the "window" advance rule.

The changed configuration field is `rotor_param`, which also changes the rotor projection: four outputs per head with bias for normalized quaternions, three without bias for Euler/exponential parameterizations. Non-embedding counts are 2,101,766 versus 2,035,206. All initialize at nominal identity, but parameterization-dependent random-number consumption means identical seed numbers do not pair the remaining initial weights. The comparison therefore does not isolate coordinate geometry from these implementation choices.

That the choice of rotation parameterization affects gradient-based learning at fixed expressivity is documented for rotation regression (Geist et al., 2024) and for fixed parameterizations of manifolds (Lezcano-Casado, 2019); those settings differ from a recurrent transition, and we do not claim that the ordering observed here transfers to them. An earlier version of this experiment used five seeds and mixed two advance rules; it is superseded by these runs, and its apparent reliability ordering did not survive the larger sample.

B.4.1 What the probes establish

The probes do not explain the learning-time gap. They measure ideal parameterizations, not a controlled intervention on the trained models; projection, initialization and normalization differences remain.

Jacobian conditioning. For each parameterization we sampled parameters at several scales and computed the singular values of the Jacobian of (parameters $\to$ rotor), restricted to the directions that move the rotor. Median and worst-case condition numbers over 200 draws:

parameter scale	normalized (med / worst)	exp (med / worst)	Euler (med / worst)
0.1	**1.00 / 1.0**	1.00 / 1.0	1.13 / 1.9
0.5	**1.00 / 1.0**	1.11 / 1.8	1.96 / 82.9
1.0	**1.00 / 1.0**	1.57 / 116.5	2.61 / 125.8
2.0	**1.00 / 1.0**	5.38 / 667.3	2.09 / 741.7

Away from zero, ideal quaternion normalization has condition number one on directions tangent to the sphere. The angle-based parameterizations can have poorly conditioned Jacobians away from identity. This provides a possible source of optimization differences, but does not measure the parameter regions visited during training.

Fitting a single rotation. Fitting parameters to match a random target rotor with Adam (tolerance 10^{-3}, 40 seeds) takes a median of 64, 55 and 66 steps for normalized, exp and Euler respectively, with no failures. Representing one rotation is equally easy in all three parameterizations.

Fitting an ordered product. Fitting 32 rotors whose ordered product matches a target takes a median of 19, 28 and 28 steps for normalized, exponential and Euler parameterizations.

The ordered-product probe has a similar ranking to training (with an exponential and Euler tie), but its approximately $1.5\times$ difference is much smaller than the successful-run solve-time ratio. These probes therefore do not explain the training effect, which also includes projection, initialization and normalization differences.

B.5 Learned transition maps

For a one-layer model with input-dependent rotors, the rotor is a function of the token alone—we checked this numerically, presenting every group element at many positions moves the induced rotation matrix by at most 1.2×10^{-7}. Each head h gives a token-to-rotor map into $\mathrm{Spin}(n)$; the diagnostics use its induced matrices $Q_h(g) \in \mathrm{SO}(n)$, discarding the lift's sign. The task composes the running product on the right, $g \leftarrow g\,x_t$, while the scan composes new transitions on the left, so for a per-head representation aligned with the task's right-multiplication convention the relevant composition law is the anti-homomorphism order, $Q_h(g_1 g_2) = Q_h(g_2)Q_h(g_1)$. Per head we therefore measure the *defect* in both orders: the mean Riemannian distance on $\mathrm{SO}(n)$ between $Q_h(g_1g_2)$ and $Q_h(g_2)Q_h(g_1)$ (task-aligned, anti order) or $Q_h(g_1)Q_h(g_2)$ (homomorphism order) over all $|H|^2$ pairs (0 = exact in exact arithmetic; independent uniform rotations give $\approx$2.2 on $\mathrm{SO}(3)$ and $\approx$2.9 on $\mathrm{SO}(5)$ on average; the two orders agree when the head's image is abelian). We also report $|\mathrm{image}|$, a greedy cluster count with tolerance 0.1 radians, and *coset collapse*, the spread within V_4-cosets relative to the overall spread (ratio below 0.15). Exact homomorphism and anti-homomorphism controls (permutation representations) read 0 in their own order and 3.6, 5.2, 9.1 and 13.7 in the other for A_4, S_4, A_5 and S_5. These finite-precision diagnostics do not certify faithfulness (`results/mechanism/head_homomorphism_orders.txt`).

Table 12: Rotor heads of trained models, all 192 heads scored; defects in the task-aligned (anti) order, with the homomorphism-order count in parentheses. The A_4 and S_4 rows diagnose the full affine checkpoints; the A_5 and S_5 rows diagnose the write-free decayed ablations. The A_4 model, which solves the training length and extrapolates imperfectly without meeting the all-lengths criterion, contains abelian-image heads with small defects in both orders; the write-free A_5 model contains four heads with mean task-aligned defects below 0.005 and 60 distinguishable images, invisible in the homomorphism order; the stalled S_4 model has none below 0.01, and its closest heads (0.014–0.033) have six distinct images. These finite-precision diagnostics do not establish exact maps.

checkpoint	outcome	anti defect < 0.01 (hom.)	min defect	$\lvert\mathrm{image}\rvert$, best heads	coset collapse
A_4	training length	$9/192$ (9)	0.0001	$3, 3, 3, 3$ (of 12)	$11/192$
A_5	solved	$4/192$ (0)	0.0010	$60, 60, 60, 60$ (of 60)	n/a
S_4	stalled	$\mathbf{0/192}$ (0)	0.0142	$6, 6, 6, 4$ (of 24)	$11/192$
S_5	stalled	$0/192$ (0)	0.0540	$1, 2, 4, 8$ (of 120)	n/a

These diagnostics do not identify the learned mechanism. The same $11/192$ heads meet the coset-collapse criterion in the A_4 (training length) and stalled S_4 models; that count alone does not explain the stall. The best A_4 heads have three measured image clusters, consistent with approximate C_3-quotient structure, but neither exact relations nor an exact kernel have been established. In the write-free A_5 model, four heads have mean task-aligned composition defects below 0.005 radians and 60 distinguishable token images, consistent with an approximate faithful A_5 representation in the induced $\mathrm{SO}(3)$ action; this does not establish exact relations, a finite reachable group, or use by the readout. The stalled S_4 model's three closest heads have six distinct images, consistent with an S_3 quotient but not an exact one. An earlier version of this table scored the homomorphism order only and missed the A_5 heads. Other heads, writes, decay, and input-dependent readout may contribute to the solution. Corollary 3 does not resolve that empirical question. These are end-of-training measurements from one checkpoint per group; when the small defects emerge is unmeasured.

B.6 Removing and changing the additive write

Four runs under the A_5/S_5 protocol of Appendix B.1 ($\mathrm{Cl}(3,0)$, 192 heads, one layer, d_{model} 384), differing only in the write (Table 13; head diagnostics of the write-free checkpoints are in Appendix B.5). `--no_write` sets $U_t \equiv 0$, giving the write-free recurrence $S_t = \lambda_t R_t S_{t-1}$; the learned initial state seeds it. The flag does not set $\lambda_t = 1$: the implementation retains learned sigmoid decay and an input-dependent query and output gate. Thus this is not the pure compact-group recurrence of Corollary 3, and the experiment does not establish uniform separation at all lengths.

Table 13: Write-free and write-rule ablations. The conditional obstruction requires hypotheses not certified by these experiments; "scope" does not predict their outcomes.

task	write	scope	outcome	acc@128/256/512	final loss
A_5	none ($U_t \equiv 0$)	decay outside pure dynamics	solved, 3,000	1.000/1.000/1.000	0.0000
A_5	convex $(1-\lambda_t)k_t v_t$	affine; outside bound	solved, 3,000	1.000/1.000/1.000	0.0000
A_5	outer $k_t v_t$	affine; outside bound	solved, 3,000	1.000/1.000/1.000	0.0000
S_5	none ($U_t \equiv 0$)	decay outside pure dynamics	stalls, 30,000	0.048/0.031/0.017	3.967

All three A_5 variants first satisfy the solve criterion at the same evaluation step (3,000), and removing the write does not move the measured S_5 plateau, which resembles that of the write-carrying arms of Table 5. These are budget-dependent empirical observations, not verification of an all-length impossibility theorem or proof that writes never matter.

B.7 Partial solutions on S_5

Reliability of the expanded families. Only one of five learned-parity $\mathrm{Pin}(4)$ runs reaches the reported plateau. In the other four, the odd-transition fraction falls from 0.50 to 0.19–0.26 within 3k steps and loss remains near 4.71. The rotor families reach their plateau in 5/5 runs. Table 5 identifies the selected plateau-reaching versor run; it is not a typical result. Gate collapse cannot alone explain failure, because the run without collapse also does not solve S_5.

For $\mathrm{Cl}(5,0)$, the step-25k checkpoint has measured sign accuracy 1.000 and within-coset entropy 3.96 nats (uniform over 60 labels: 4.09). With the $\mathrm{SO}(5)$ Riemannian metric, no head has mean composition defect below 0.3 radians in either order (minimum 0.77, median 2.82, against 2.94 for independent uniform $\mathrm{SO}(5)$ rotations), and the two closest heads have 120 distinct images. A nearest-centroid decoder of composed rotations reaches accuracies 1.000 and 0.372 at lengths one and two, then chance at lengths four and above. An exact homomorphism control validates the decoder at 1.000 for every tested length. Accumulation factors 1, 4 and 8 give identical measured loss and accuracy in the numerical check.

The parity-gate control is confounded. Our pre-registered control deleted the gate, fixing half the heads odd and half even. It does not solve either (loss 4.71 against the random-label reference 4.79), and does not show the same sign-classification result. This changes the transition family as well as removing the learned gate: an odd one-step transition cannot be the identity, although products of odd transitions can. It is not a clean isolation of gate optimization, and we report it as inconclusive.

What the sign probe establishes. We measure whether the output identifies permutation parity despite low group-element accuracy. Across checkpoints separated by 500 training steps, sign accuracy rises from 0.51 to 1.000 and probability on the correct parity coset from 0.51 to 0.999, while within-coset entropy remains 4.030 nats. This transition appears in the probed $\mathrm{Cl}(3,0)$ and $\mathrm{Cl}(4,0)$ runs at different training times (1.5k and 17k steps). For reference, uniform guessing within the correct coset gives loss $\ln 60 = 4.09$ and group accuracy $1/60 = 0.017$.

The models do not learn an exact sign tracker at every length. Across three runs, sign accuracy ranges from 0.92 to 1.00 at twice the training length and from 0.47 to 0.70 at four times it; two seeds of the same configuration give 0.998 and 0.923 at twice the length. Moreover, the measured loss and group accuracy differ from the ideal sign-only reference. The probe shows a partial output solution on sampled words, not that sign is the only information in the state or that the state realizes an exact quotient.

C Language-memory scope and controls

These experiments test the scope of compact rotor memory, not the loss theorem. The standalone comparisons measure language-modelling performance; the hybrid comparison and paired identity intervention test different explanations of the hybrid gain. Their protocols and evaluations are not interchangeable.

C.1 Standalone language modelling at 125M parameters

The compact rotor state does not match the associative-memory baselines under this protocol (Table 14, Figure 2). The state counts below measure stored payload, not hardware efficiency. A separate WikiText-103 comparison of two rotor parameterizations is in Appendix C.2.

Protocol.

- **Data & Architecture:** FineWeb-Edu with the GPT-2 tokenizer. Twelve pre-norm blocks with SwiGLU of hidden size $2\times$.
- **Training:** Sequence length 512, batch 8 with 4-step gradient accumulation (16,384 tokens per step), 61,000 steps (1.0B tokens). Learning rate 6×10^{-4} with 6,100 warm-up steps and cosine decay to 10%, weight decay 0.1 on matrices, bf16, one seed per model. Optimizer settings are identical across architectures and untuned.
- **Dimensions:** Widths are set per model so that non-embedding parameters match to within 5%: d_{model} 1280 for the rotor model (32 heads; 1088 at 170 heads), 1024 for the Transformer, GLA, and DeltaNet, 896 for Mamba-2, and 640 for Gated DeltaNet. Baselines use the `flash-linear-attention` layers dropped into our block.
- **Evaluation:** Reported perplexity is on a fixed held-out sample of 500 batches (2.05M tokens), evaluated identically for every checkpoint.
- **Variant Runs:** The 50M pair uses the same recipe at d_{model} 768 (rotor) and 640 (GLA), stopped at step 20,000. Mamba-2 and the 170-head run were trained on a rented RTX 5090; only perplexity is compared across machines.

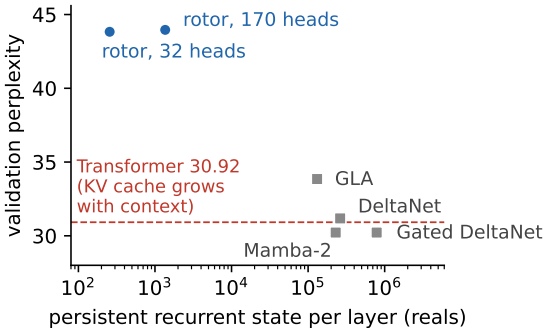


Figure 2: At 125M parameters, perplexity against persistent recurrent state per layer. The rotor carries two to four orders of magnitude less state and pays for it; widening it $5.3\times$ (32 to 170 heads) moves perplexity 0.3%. That intervention does not close the gap, without excluding other capacity or optimization explanations. This motivates testing a separate associative memory (§4.4).

Table 14: FineWeb-Edu, approximately 125M non-embedding parameters, 1B tokens, matched schedule, trained at sequence length 512; one seed per row. Perplexity on 2.05M held-out tokens at the training length and at 2,048 tokens without retraining. Recurrent state is per token per layer, in reals; a Transformer's grows with context. Counts are recurrent payload, excluding auxiliary convolution caches (Appendix D.3). †2,048-token evaluation unavailable: the evaluator retained a 512-position RoPE cache. ‡kernel fault on our hardware at 2,048.

model	ppl @512	ppl @2048	recurrent state
Transformer (SDPA)	30.92	—†	grows with context
GLA	33.85	31.74	131,072
Mamba-2	30.22	—‡	229,376
DeltaNet	31.18	30.26	262,144
Gated DeltaNet	30.22	28.07	786,432
Rotor recurrence, $\mathrm{Cl}(3,0)$, 32 heads (ours)	43.83	42.38	**256**
Rotor recurrence, $\mathrm{Cl}(3,0)$, 170 heads	43.96	—	1,360

C.2 WikiText parameterization comparison

This comparison tests two implemented rotor parameterizations in language modelling, under a separate protocol from the 125M study.

WikiText-103 with the GPT-2 BPE tokenizer (vocabulary 50,304), 117.9M training tokens. Twelve pre-norm blocks, d_{model}=768, SwiGLU with hidden size $2\times$, sequence length 512, batch 8 with 4-step gradient accumulation (16,384 tokens per step), 9,000 steps (≈147M tokens, 1.25 epochs), learning rate 6×10^{-4} with 900 warm-up steps and cosine decay to 10%, weight decay 0.1 on matrices only, bf16 autocast. Rotor layers use $\mathrm{Cl}(3,0)$ with 32 heads, data-dependent rotors, ℓ_2-normalised q and k; 48.4M non-embedding parameters (48.1M for the Euler parameterization, which has three rotor outputs per head instead of four). Validation perplexity on 20 batches of 8 sequences every 500 steps; the two runs are step-aligned and compared at equal steps (Figure 3).

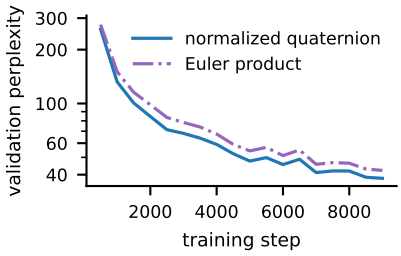


Figure 3: WikiText-103, 48M non-embedding parameters, the same recipe with different implemented rotor parameterizations (including projection and initialization differences). Validation perplexity at matched steps; final 38.2 (quaternion) vs 42.3 (Euler), a 10.7% gap.

C.3 Hybrid and constant-rotation controls (§4.4)

All 50M arms use FineWeb-Edu, sequence length 512, 16,384 tokens per optimizer step, learning rate 6×10^{-4}, 6,100 warm-up steps and a 61k-step cosine horizon, stopping at step 20k (327.68M tokens). The twelve-layer hybrids alternate six rotor and six memory layers at width 704; DeltaNet and Transformer controls have twelve memory layers at width 640. Rotor heads use $n = 3$, 32 heads, normalized parameters, normalized q, k, and additive writes. DeltaNet uses eight heads; attention uses eight heads. The DeltaNet comparison has seeds 0 and 1; the constant-rotation and attention comparisons have seed 0 only.

The constant-rotation control replaces the input-to-rotor projection with learned per-head parameters (`--rotor learned`). It is constant in token/time, not in training. The DeltaNet hybrid has 50,386,082 non-embedding parameters, the constant version 49,845,410, and the control 49,322,560. The analogous attention counts are 50,301,074, 49,760,402 and 49,168,000. Both constant variants have about 1.1% fewer parameters than their token-dependent counterparts.

Final evaluations use 500 batches (2.05M tokens); disjoint 250-batch halves are reported separately. The evaluation loader seeds from the training seed, so comparisons across arms are paired within a seed but absolute perplexities across seeds use different windows. For the attention trio, perplexities are 51.17, 50.58 and 50.66 (control, token-dependent, constant); the constant and token-dependent hybrids beat the control at 11/20 and 14/20 in-training checkpoints.

50M state accounting. Counts follow the saved run configurations and checkpoint projection shapes. Rotor-only payload is $12 \cdot 32 \cdot 8 = 3{,}072$ scalars; the hybrids have six rotor layers, hence $1{,}536$. DeltaNet's eight heads have key/value widths 80 in the control and 88 in the hybrid, giving $12 \cdot 8 \cdot 80^2 = 614{,}400$ and $6 \cdot 8 \cdot 88^2 = 371{,}712$ matrix-state scalars. The GLA control has four heads of key/value widths 80/160 and twelve layers, also $614{,}400$. DeltaNet's three length-four convolution buffers add $12 \cdot 3 \cdot 640 \cdot 4 = 92{,}160$ and $6 \cdot 3 \cdot 704 \cdot 4 = 50{,}688$ scalars. Table 15 separates these auxiliary caches from the rotor/matrix payload. It makes no measured bandwidth claim.

Results. Paired differences on disjoint evaluation halves (Table 15, Figure 4) are 2.53/2.57 and 2.15/2.26; these check sensitivity to the sampled text, not training variance. Beside attention the control, token-dependent and constant hybrids reach 51.17, 50.58 and 50.66 (one seed).

Table 15: FineWeb-Edu, approximately 50M non-embedding parameters, 20k steps (328M tokens). Perplexities use 2.05M-token evaluations; compare within each seed, because seed 1 uses a different held-out window. State is the total stored rotor-plus-matrix payload in scalars, excluding short-convolution caches and attention's KV cache (convolution caches in Appendix C.3). Constant means learned but token-independent, not frozen during training. A dash denotes an unrun seed.

model	params (M)	ppl, seed 0	ppl, seed 1	rotor + matrix state
Rotor only	48.39	74.34	—	3,072 + 0
GLA	49.36	57.67	—	0 + 614,400
DeltaNet control	49.32	53.43	52.43	0 + 614,400
Rotor + DeltaNet	50.39	**50.88**	**50.22**	1,536 + 371,712
constant rotation	49.85	50.81	—	1,536 + 371,712
Transformer	49.17	51.17	—	0 + KV cache
Rotor + Transformer	50.30	50.58	—	1,536 + KV cache
constant rotation	49.76	50.66	—	1,536 + KV cache

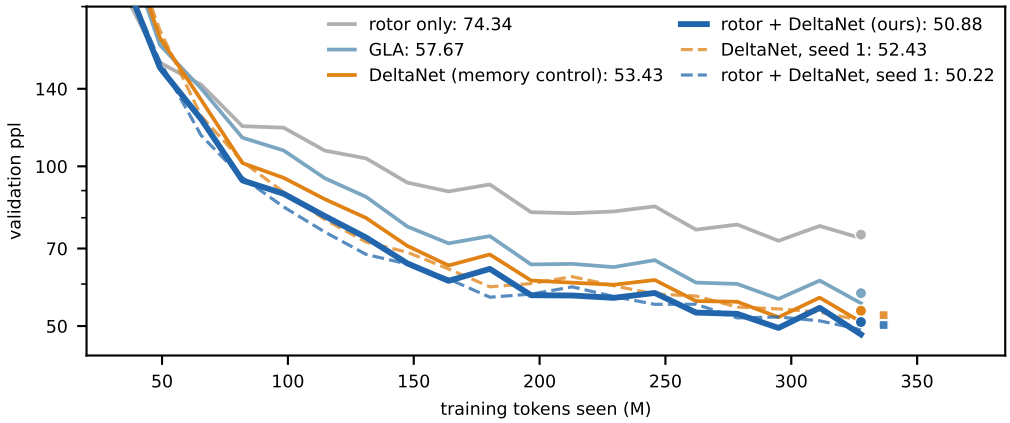


Figure 4: Matched-token training curves, 50M comparison. Lines: in-training evaluations; markers: final 500-batch evaluations on a different sample, with values in the legend (circles seed 0; squares seed 1, offset right for legibility). The disjoint-half ranges, not confidence intervals, are narrower than the markers and are not drawn: DeltaNet 53.32–53.54 and hybrid 50.75–51.01 (seed 0); 51.87–52.99 and 49.72–50.73 (seed 1). Solid lines are seed 0; dashed lines are seed 1, run for the DeltaNet control and the hybrid only.

C.4 Paired identity/constant intervention

The protocol was frozen before research training: one paired seed (300), 20,000 successful updates per arm, no efficacy stopping, replacement seeds or rescue tuning. Both models alternate six rotor and six DeltaNet layers, width 704, 32 rotor heads, eight DeltaNet heads, vocabulary 50,304, bf16 autocast and fp32 parameters. Both start at identity rotations with identical full state dictionaries, including all non-rotor weights. Named independent initialization and training RNG streams give paired batches. Identity coefficients are fixed buffers; the constant arm learns 768 quaternion coefficients. Both execute the same rotational scan and have the same 373,248-scalar recurrent payload plus 50,688 auxiliary convolution scalars.

Unlike the randomly initialized constant control of the 50M comparison, both arms use $q/\max(\|q\|, 10^{-8})$; nonfinite parameters or raw norm $\leq 10^{-6}$ abort the run. This removes additive-epsilon contraction, not finite-precision error. Training uses FineWeb-Edu, length 512, batch eight, accumulation four (16,384 tokens/update), AdamW $(\beta_1, \beta_2) = (0.9, 0.95)$, epsilon 10^{-8}, matrix-only weight decay 0.1 and gradient clipping at 1. The learning rate follows the 50M comparison's

61k-step cosine horizon, with 6,100 warm-up steps, peak 6×10^{-4} and floor 10%; stopping at 20k does not compress that horizon. Each arm consumes 327,680,000 tokens.

Validation uses 20 fixed batches every 1k updates from the first half of the held-out token array. Final evaluation occurs only after completion: 500 fixed batches, 4,000 windows, 2,048,000 token predictions, from its disjoint second half. Windows never cross the split boundary but may overlap within each half; this is not a new document-deduplication audit. Both arms use the same registered windows. The primary contrast is final token-mean NLL(identity) minus NLL(constant), positive favoring learned transport (Table 16). No training-variance interval or equivalence test is available.

Table 16: Paired intervention (seed 300), final held-out evaluation. Lower is better. These are not directly comparable to the 50M comparison endpoints in Table 15: initialization, normalization and evaluation windows differ.

rotation	non-embedding parameters	NLL (nats/token)	perplexity
Identity (fixed)	49,844,642	3.951727	52.025155
Constant (learned)	49,845,410	3.953977	52.142332

Both arms completed all 20,000 attempts successfully, without skipped updates or retries. The contrast is -0.002249789 nats/token: identity's perplexity is about 0.225% lower. The pre-specified 0.01-nat practical target was descriptive, not an equivalence margin or an inferential threshold. No token-dependent or pure-DeltaNet arm was trained in this intervention; it neither retroactively deconfounds the 50M hybrid comparison nor measures the gain from interleaving generic layers.

Numerical checks preceded training. The archived evidence includes the frozen manifest, update journals, evaluation losses and source/data/checkpoint audit receipts. These verify recorded provenance and arithmetic, not an independent training reproduction. Launch history and portable audit instructions are in supplementary `results/identity_transport_20260925/`; checkpoints and data are excluded.

D Implementation and state accounting

This section defines the algebra conventions used by the implementation, explains the state update, and separates stored-state counts from execution costs.

D.1 Clifford algebra conventions

$\mathrm{Cl}(n,0)$ is the real associative algebra generated by $e_1, \ldots, e_n$ subject to $e_i^2 = 1$ and $e_i e_j = -e_j e_i$ for $i \neq j$. A basis is the set of *blades* $e_{i_1} e_{i_2} \cdots e_{i_k}$ with $i_1 < \cdots < i_k$, one for each subset of $\{1, \ldots, n\}$, so $\dim \mathrm{Cl}(n,0) = 2^n$; the number k is the *grade*. A general element (a *multivector*) is a real combination of blades; $\langle A \rangle_k$ denotes its grade-k part. We index blades by bitmask, so for $n = 3$ the basis in order is 1, e_1, e_2, e_{12}, e_3, e_{13}, e_{23}, e_{123}.

The product. Concatenate blades and reorder their generators: swapping distinct generators changes the sign, and $e_i^2 = 1$ cancels repeated pairs. For two vectors $a = \sum a_i e_i$, $b = \sum b_i e_i$,

$$ab \;=\; \underbrace{\textstyle\sum_i a_i b_i}_{a \cdot b \text{ (grade 0)}} \;+\; \underbrace{\textstyle\sum_{i<j} (a_i b_j - a_j b_i)\, e_i e_j}_{a \wedge b \text{ (grade 2)}}.$$

The implementation uses the dense contraction $(AB)_k = \sum_{i,j} A_i B_j C_{ijk}$ with a fixed structure tensor $C \in \{-1, 0, 1\}^{2^n \times 2^n \times 2^n}$.

Reversion and norm. Reversion $\tilde{A}$ reverses the order of generators in every blade, which multiplies a grade-k blade by $(-1)^{k(k-1)/2}$. For an even element in dimensions at most three, $A\tilde{A}$ is the scalar $\|A\|^2$. In dimension four it may also have a pseudoscalar component; in higher dimensions additional grades can occur.

Even subalgebra, rotors, versors. The even-grade elements close under the product and form the *even subalgebra* $\mathrm{Cl}(n,0)^+$ of dimension 2^{n-1}. For $n=2$ it is spanned by $1, e_{12}$ with $e_{12}^2=-1$, i.e. $\mathbb{C}$; for $n=3$ it is spanned by $1, e_{12}, e_{13}, e_{23}$ and is the quaternions $\mathbb{H}$; for $n=4$ the pseudoscalar $\omega = e_{1234}$ is central in $\mathrm{Cl}(4,0)^+$ with $\omega^2=+1$, and the idempotents $(1\pm\omega)/2$ split it as $\mathbb{H}\oplus\mathbb{H}$. A *rotor* is an even product of unit vectors, hence satisfies $R\tilde{R}=1$ and preserves the vector subspace under conjugation; norm normalization alone is not a general definition in arbitrary dimension. It acts on a vector v by $v \mapsto Rv\tilde{R}$, which is a proper rotation, and the rotors form $\mathrm{Spin}(n)$, a double cover of $\mathrm{SO}(n)$ (R and $-R$ give the same rotation). A unit vector u acts by $v \mapsto -uvu$, a reflection in the plane orthogonal to u; products of unit vectors are *versors*, even ones are rotors, odd ones realise improper rotations, and together they form $\mathrm{Pin}(n)$, a double cover of $\mathrm{O}(n)$. Everything in §2 uses only these facts; the reference is Lounesto (2001).

D.2 Propagation of past writes

To compare how rotor and DeltaNet states retain past information, consider two states receiving the same future inputs. Ideal unit-rotor updates give

$$\Delta S_t = \lambda_t R_t \Delta S_{t-1}, \qquad \|\Delta S_t\| = \lambda_t \|\Delta S_{t-1}\|.$$

All directions contract by the same scalar. By contrast, DeltaNet's update $M^+ = M(I-\beta kk^\top) + \beta vk^\top$, with unit key k, gives

$$\|\Delta M^+\|_F^2 = \|\Delta M\|_F^2 - \beta(2-\beta)\|\Delta M k\|^2.$$

For $0<\beta<2$ it contracts the key-addressed component. This difference motivates the hybrid comparison but does not establish its causal explanation.

For invertible rotor transitions, a change of coordinates also clarifies their role. With $P_t = R_t \cdots R_1$ and $Z_t = P_t^{-1} S_t$,

$$Z_t = \lambda_t Z_{t-1} + P_t^{-1} U_t.$$

The rotation remains in the transformed writes and readout; this identity does not remove it from the full model. If $R_t = R$, rotational transport depends on lag as R^{t-s}, while decay, writes and readout may still depend on the input. The epsilon normalization of the group and 50M runs adds contraction and can produce a zero transition, for which the inverse-coordinate expression is undefined.

D.3 Stored state and analytic access counts

We distinguish stored state from physical memory traffic. Table 17 counts one read and one write per recurrent-state scalar, or attention's cache append and read, for the 125M configurations. It excludes projections, temporary storage, convolution buffers and transfers between memory levels. Cache residency and kernel fusion determine actual device-memory traffic; these counts are not bandwidth or runtime measurements.

The rotor payload is 3,072 scalars across twelve layers (6 KiB if stored in bf16). The matrix-state counts follow the saved head dimensions: GLA 4 heads with key/value widths 128/256; DeltaNet 4 with 256/256; Gated DeltaNet 6 with 256/512; Mamba-2 28 with dimension 64 and SSM state size 128. Excluded length-four convolution caches follow the same head dimensions: DeltaNet 12,288 scalars per layer (three buffers of width $4\times256 = 1{,}024$), Gated DeltaNet 24,576 (query and key buffers of width $6\times256 = 1{,}536$ and a value buffer of width $6\times512 = 3{,}072$) and Mamba-2 8,192 (one buffer over the expanded input of width $28\times64 = 1{,}792$ plus two state blocks of 128). Appendix C.3 separately accounts for the 50M configurations.

A direct quaternion product uses 16 scalar multiplications, compared with 64 for two dense 4×4 left-action matrices. This algebraic example is not the operation count of the unfused implementation or an optimized baseline scan. Matched parameter counts do not imply matched FLOPs or execution costs.

D.4 Numerical implementation and hardware

Most runs use two AMD Radeon AI PRO R9700 (32 GB, ROCm 7.2, PyTorch 2.11); the Mamba-2 and widened 125M rotor runs used a rented RTX 5090 as noted above. The scan is unfused

Table 17: Recurrent payload and analytic read/write counts, 12 layers, 125M configurations; numbers are scalars, not measured hardware transfers. Auxiliary caches are excluded. Attention's KV cache is shown at context 2,048.

model	state per layer	state, all layers	read+write per token
Rotor recurrence, $\mathrm{Cl}(3,0)$, 32 heads	256	3,072	6,144
GLA	131,072	1,572,864	3,145,728
Mamba-2	229,376	2,752,512	5,505,024
DeltaNet	262,144	3,145,728	6,291,456
Gated DeltaNet	786,432	9,437,184	18,874,368
Transformer KV cache (ctx 2,048)	grows	50,331,648	50,356,224

PyTorch. The geometric product uses blocks of at most 2^{19} rows to avoid a driver fault and was checked against CPU to 10^{-7} at all trained sizes; the supplementary `docs/ROCM_MATMUL_BUG.md` records the fault and reproducer. A step whose loss or gradient norm is non-finite is skipped and counted in the group and 50M trainers; the paired-intervention trainer instead aborts. Every reported run has zero such steps. Seeds are specified per experiment (0–9 in the ten-seed parameterization study); we report every run we completed under the stated protocol, including negatives.