RESIDUAL STREAMS READ, RECURRENT STATES REMEMBER: THE GLOBAL WORKSPACE IN MAMBA MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can the global-workspace account of transformer representations extend to state-space language models? We fit Jacobian lenses to the residual streams and recurrent states of Mamba-1, Mamba-2 and Mamba-3, using the original 1000-prompt recipe. Joint residual–state readouts improve recovery of known intermediate concepts over the residual lens on at least five of six task families in every tested Mamba checkpoint. On Mamba-2, state alone exceeds residual and logit lenses on all six families; a normalised joint readout improves on both components on five. Temporal maps and word-list experiments show earlier content remaining state-readable as residual visibility changes. We also propose sign-guarded steering, which improves target top-five success over coordinate exchange on matched verbal-report trials in five models. Recurrent state alone supports this verbal access. These gains do not extend consistently to relational answers: guarded edits often output the edited concept itself, and Mamba-3's joint edits can disrupt successful state-only redirection. Recurrent state thus provides a complementary carrier of workspace content, whose recovery, persistence and causal uses require separate measurements.

1 INTRODUCTION

Consider "The US state where the Space Needle is located shares its northern border with _". A model can identify the intermediate city, *Seattle*, before producing the answer, *Canada*. Where does this intermediate remain available as the computation proceeds? Global workspace theories address a related coordination problem in cognition: selected information becomes broadly accessible to specialised systems (Baars, 1988; Dehaene & Naccache, 2001).

Gurnee et al. (2026) investigate this account in transformer language models. Their Jacobian lens (J-lens) averages input–output derivatives to identify token-labelled directions associated with future verbalisation. The resulting J-space supports intermediate-concept recovery and tests of verbal report, reasoning, flexible generalisation and other workspace properties. Extending this analysis matters as language models increasingly use recurrent computation. Looped models such as Ouro repeatedly update latent representations (Zhu et al., 2025); Jacobian-lens experiments on looped transformers find recurrence-dependent read and intervention horizons (Wang & Reid, 2026).

Selective state-space models introduce recurrence across *tokens*, carrying a compressed prefix in an SSM state (Gu & Dao, 2023; Dao & Gu, 2024; Lahoti et al., 2026). This gives a second observation point alongside the residual stream, including in hybrids such as Nemotron 3.5 Lightning (NVIDIA Corporation, 2026). We extend the lens to that recurrent state and ask where intermediate concepts can be recovered, how they persist, and whether edits can make them available for verbal report and relational answers.

Our contributions follow these three questions. First, we introduce a full-path state lens and a readout-only control, mapping state into the same final-residual coordinates as the residual lens. Second, we show that their joint readouts improve recovery across Mamba checkpoints and that recurrent state exposes earlier content across later tokens. Third, we compare a full suite of workspace properties and propose sign-guarded steering, which improves verbal report over coordinate exchange on matched

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.

trials. Its downstream effects depend on the task and intervention site. Observing both representations and editing both representations therefore have different consequences.

2 State and joint Jacobian lenses

2.1 Reading residual streams and recurrent states

Let $h_{\ell,t}$ be the residual activation at source block ℓ and token t, and $s_{\ell,t}$ its flattened post-update SSM state. Schematically, $s_{\ell,t} = A_{\ell,t}s_{\ell,t-1} + B_{\ell,t}x_{\ell,t}$ and $y_{\ell,t} = C_{\ell,t}s_{\ell,t} + D_{\ell,t}x_{\ell,t}$, where y is the scan output before gating and projection. Let z_u be the final-block residual before final normalisation, and $Z(x) = \sum_{u \in V_x} z_u$ sum its valid fitting positions. We fit

$$J_\ell^h = \mathbb{E}_{x,t}\left[\frac{\partial Z(x)}{\partial h_{\ell,t}}\right], \tag{1}$$

$$J_\ell^{s,\mathrm{full}} = \mathbb{E}_{x,t}\left[\frac{\partial Z(x)}{\partial s_{\ell,t}}\right], \tag{2}$$

$$J_\ell^{s,\mathrm{read}} = \mathbb{E}_{x,t}\left[\frac{\partial Z(x)}{\partial y_{\ell,t}} C_{\ell,t}\right]. \tag{3}$$

The full-path state lens includes paths through both the current output and later states of the source block. In the readout-only control, $C_{\ell,t} = \partial y_{\ell,t}/\partial s_{\ell,t}$ is the linear state readout at the captured inputs, and only the path through $y_{\ell,t}$ is retained. Its downstream derivative includes gating, output projection and all later blocks. Thus J_{read} excludes propagation through future states of the *source block*, while retaining paths to future target positions through downstream blocks. Both variants are $d_{\mathrm{model}} \times d_s$ maps into final residual coordinates; their transported states are decoded with the model's final normalisation and unembedding. Fitting averages derivatives of the frozen model without training a decoder on concept labels.

Writing $r_h = J_\ell^h h_{\ell,t}$ and $r_s = J_\ell^s s_{\ell,t}$, we combine the transported representations without an additional fit:

$$r_{\mathrm{sum}} = r_h + r_s, \qquad r_{\mathrm{nsum}} = \frac{r_h}{\|r_h\|_2} + \frac{r_s}{\|r_s\|_2}. \tag{4}$$

Main results use post-update state. A pre-update convention gives the same qualitative result except on typo correction, where the readout token itself writes the correction into state. A fixed state-coordinate permutation controls for dimensionality while destroying the fitted correspondence. For Mamba-3, state and derivatives use matching current-token rotary coordinates. Derivative recurrences and numerical checks are in Appendix I.1.

For a token v with unembedding column u_v, a lens direction is proportional to $J_\ell^\top u_v$. Residual directions have d_{model} coordinates and state directions d_s. Shared output coordinates therefore do not imply shared source dictionaries. We test the lenses' output-side geometric overlap in Appendix H.

2.2 Fitting and evaluation

We study Mamba-1 2.8B, Mamba-2 2.7B and Mamba-3 SISO/MIMO 1.5B, with Pythia-2.8B (Biderman et al., 2023) and Nemotron-3.5 as transformer and hybrid comparisons. State readouts cover all 63 source blocks of Mamba-1/2 and all 23 of Mamba-3; the final block is the fitting target. MIMO has readout results but no full state-intervention suite. Nemotron's joint readout omits its attention KV cache. Checkpoints differ in training, scale, tokenisation and competence; our primary comparisons are within models (Table 4).

All paper-comparable lenses use 1000 WikiText-103 prompts truncated to 128 tokens. Both source and target positions use the reference mask $V_x = \{16, \ldots, T_x - 2\}$, with zero-based indices; excluded tokens remain in the forward computation. For N_{p} prompts and any derivative matrix $M_\ell(x,t)$ in Equations 1–3, we estimate

$$\mathbb{E}_{x,t}[M_\ell(x,t)] \equiv \frac{1}{N_{\mathrm{p}}} \sum_{x=1}^{N_{\mathrm{p}}} \frac{1}{|V_x|} \sum_{t \in V_x} M_\ell(x,t). \tag{5}$$

Source derivatives are averaged within each prompt, then over prompts with equal weight. Target effects are summed inside $Z(x)$, including current and future positions, without division by the number of targets or causal source–target pairs. Accumulation uses fp32. This matches the reference residual-lens estimator, including for the Pythia control; shard and sensitivity details are in Appendix A.3.

For each output coordinate, we differentiate the sum of valid target activations with respect to captured scan outputs. Multiplication by the local state readout gives the readout-only derivative; a reverse recurrence adds paths through future source-block states. Equivalent weighted sums avoid materialising every state Jacobian. Appendix I.1 gives the recurrence, accumulation procedure and numerical checks.

The primary evaluation recovers known intermediate concepts on six official families: multihop, multilingual mapping, order of operations, poetry, association and typo correction. For each intermediate, we take its best vocabulary rank over fitted source blocks and integrate pass@k against $\log_{10} k$, for integer $k = 1, \ldots, 100$, to obtain normalised AUC. Items receive equal weight and only eligible single-token surfaces are scored (Appendix B). We compare per-family scores and report matched middle-half-band results alongside any-block recovery. Next-token agreement is supplementary description, not our lens-quality criterion. Persistence measurements track ranks across token positions and recover earlier words from 80-word lists using rank and sparse-pursuit readouts.

2.3 Interventions and endpoints

Let x be a residual or flattened-state column vector and a, b unit source and target lens directions. We compare signed projection transfer (N), sign-guarded steering (G) and coordinate exchange (X):

$$x_N = x + (x^\top a)(b - a), \tag{6}$$

$$x_G = x + |x^\top a|(b - a), \tag{7}$$

$$x_X = x + V(P - I)V^+ x. \tag{8}$$

Here $V = [a, b]$ and P exchanges the two coordinates. Our primary exchange uses unit columns, unweighted by the final norm's gain, with a float64 pseudoinverse; raw-column and legacy variants remain separate (Appendix C.2). Guarding orients the target-minus-source displacement, giving a non-negative local contrast change:

$$(x_G - x)^\top (b - a) = 2|x^\top a|(1 - a^\top b). \tag{9}$$

This local property does not guarantee downstream task improvement. Operators have different native edit magnitudes and trajectories.

Edits apply at every prompt position in a fixed middle-half block band. State edits use J_{full} and persist inside the scan; R, S and RS denote residual-only, state-only and simultaneous edits. Comparisons freeze within-model prompts, targets and sites, with three random-direction controls matched to each arm's realised per-site norm. Coverage differences, historical J_{read} clamps and paired simultaneous intervals are specified in Appendices C–D.

For Mamba-1, verbal edits cover all 33 state-band blocks, whereas the suite uses 12 sampled state blocks alongside all 33 residual-band blocks. Mamba-2 uses all 33 for both interfaces, Mamba-3 all 13, and Nemotron 12 state blocks alongside 27 residual blocks. Comparisons between models retain these coverage differences.

Verbal report scores whether the edited category member enters the next-token top five. Two-hop reasoning and flexible generalisation instead score its *relational answer*: after Brazil is edited to Mexico, the requested language is Spanish; after France is edited to Canada, the capital is Ottawa. The suite scores greedy output, including its saved second step after whitespace. We separately audit edited-concept output and originally correct trials. Prompted introspection uses its own direct-injection protocol.

3 RECURRENT STATES EXPOSE COMPLEMENTARY INTERMEDIATE CONTENT

State readouts improve recovery across the tested Mamba checkpoints. The best evaluated state variant exceeds the residual J-lens on five of six families for Mamba-1 and Mamba-3 SISO, six for Mamba-2 and four for Mamba-3 MIMO. Plain joint readouts improve on residual alone on all six families for Mamba-1/2 and five for both Mamba-3 variants (Table 6). These are directional comparisons, rather than significance claims for every difference.

Table 1: Mamba-2 intermediate recovery over all 63 source blocks, measured by normalised pass@k AUC (not task accuracy). The normalised joint readout combines residual and state after separately normalising their transported representations. Bold marks the best non-oracle readout in each row.

Task	Residual J^h	State J_{full}	Joint normalised	Permuted-state joint
Multihop	0.475	0.557	**0.679**	0.431
Multilingual	0.362	0.615	**0.647**	0.312
Order of ops.	0.460	0.795	**0.807**	0.515
Poetry	0.029	0.101	**0.114**	0.092
Association	0.058	**0.237**	0.231	0.117
Typo	0.303	0.706	**0.714**	0.350

Mamba-2 illustrates both the state advantage and the value of combining interfaces (Table 1). Its J_{full} lens exceeds the residual, logit and tuned lenses on all six families; additional baselines appear in Table 9. The normalised joint readout exceeds both components on five families, raising multihop AUC from 0.475 for residual and 0.557 for state to 0.679. Association is the exception: the weaker residual term slightly dilutes state recovery. The plain sum exceeds residual on every family but exceeds state on only two, so the gain depends on the combination. Poetry remains difficult, and the controls show that non-zero recovery alone is insufficient evidence of a useful reader.

Any-block recovery summarises what each interface can reveal, but does not locate its strongest readout. On multihop, multilingual and arithmetic tasks, state recovery peaks earlier than residual recovery across all four Mamba checkpoints; typo correction often reverses this ordering. The state advantage also persists within a shared range of intermediate depths: restricting both readers to the same middle half of each network preserves stronger state recovery on most families in every checkpoint (Appendix B). We next follow readability across token positions as the prompt unfolds.

4 EARLIER CONTENT REMAINS STATE-READABLE

Return to the Space Needle prompt. Both lenses recover *Seattle*, which is never named in the input, at the end of "Needle" (Figure 1). At state block 28, Seattle remains rank one through the final prompt token. In the residual maps, it falls outside the top 100 at every source block over "its northern border", then reaches rank one at a later block at the final position. State thus keeps the city readable through this residual gap.

The emerald-birthstone example gives a contrasting pattern: *May* becomes readable late in the residual stream but remains outside the state's top 100 throughout. Mamba-2 answers both prompts correctly. These selected traces illustrate complementary visibility; they do not estimate its prevalence or demonstrate a natural transfer between interfaces.

Word-list prompts provide aggregate evidence that does not depend on selecting a multihop story. As an 80-word list grows, we count earlier words appearing among the state's top-five predictions at any evaluated block. Scores average comma positions in the final quarter of the list. With lenses fitted at all 63 source blocks and readouts evaluated in the middle-half band, Mamba-1 and Mamba-2 retain respectively 7.5 and 6.7 earlier words. Non-negative pursuit on the transported state gives 18.7 and 18.6 words. The first estimator counts highly ranked vocabulary items; the second decomposes the transported representation into a sparse concept inventory. Both measure lens-recoverable content, rather than the model's total memory or its ability to recite the list.

Coverage materially affects these measurements. Expanding Mamba-1 from a sparse fitted-block subset to all 63 source blocks raises J_{full} rank-based occupancy from 0.3 to 7.5 words. Expanding

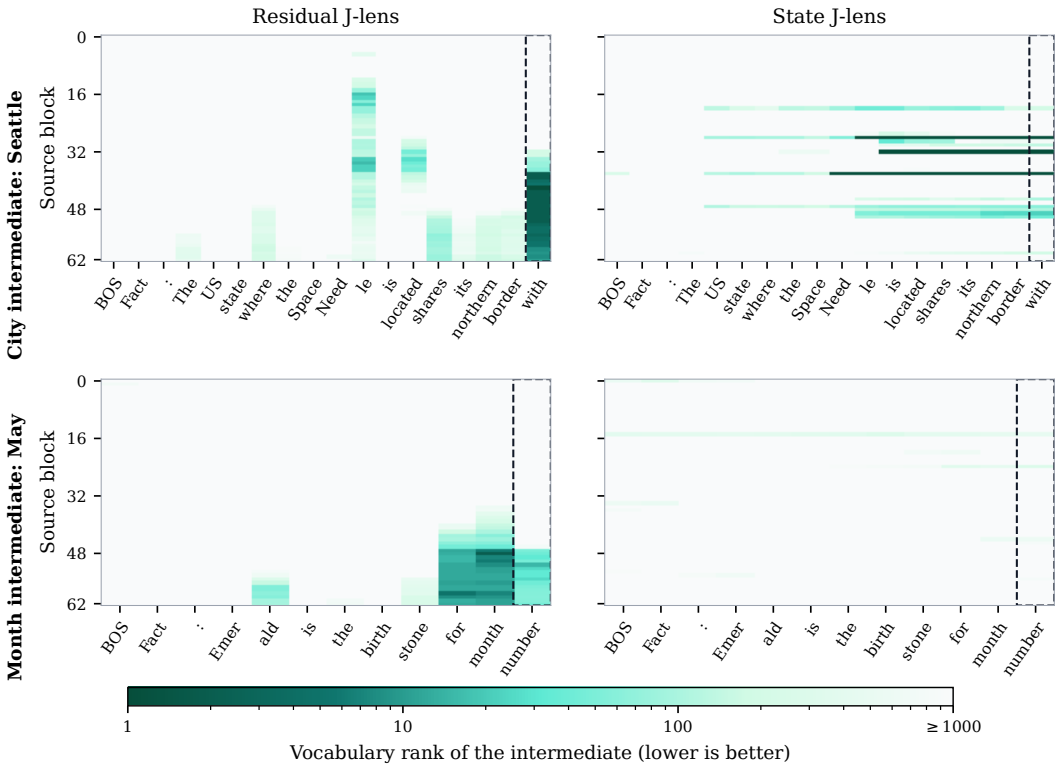

Figure 1: Selected Mamba-2 rank maps across all 63 source blocks, using the post-update convention. *Seattle* remains top-ranked at state block 28 from the end of "Needle" through the final readout, while its residual visibility varies. In the emerald question, *May* becomes readable late in the residual stream and remains outside the state's top 100 throughout. Both prompts are answered correctly.

Mamba-2's fitted coverage from 23 to 63 source blocks raises it from 3.1 to 6.7, and raises pursuit occupancy from 10.1 to 18.6. Both expanded fits cover all 33 middle-half readout blocks. Matched coverage removes an earlier apparent monotonic trend across Mamba generations: Mamba-1 and Mamba-2 have similar pursuit occupancy despite different state parameterisations. These controls show why coverage must be specified when comparing readable memory. Full conventions and a worked list appear in Appendices A.6 and E.2.

5 Workspace properties and intervention outcomes

5.1 Workspace properties across models

Table 2 brings the complete workspace suite together. Its first three rows use coordinate exchange for our measurements; Qwen values are attributed historical references under different conventions. Exchange reaches the requested two-hop answer on 5–15 of 90 trials and the flexible-generalisation answer on 12–30 of 192 at the displayed sites. These outcomes characterise a particular edit and endpoint. They do not rank architectures or establish that a representation is inaccessible by other constructions.

State-readable word capacity exceeds the transformer residual references across the recurrent models. The state-only lower bounds are 7.5, 6.7, 5.9 and 3.1 words for Mamba-1/2/3 and Nemotron, respectively; each exceeds Qwen's 0.7 and Pythia's 2.1 residual-readable words. The ranges in Table 2 bound the residual–state union between the state count and the sum of the two counts. They therefore express uncertainty about overlap, rather than confidence intervals or total memory capacity.

Table 2: Workspace properties across models. Our measurements in the first three rows use X_{unit} exchange: simultaneous residual–state edits (RS) for recurrent models and residual edits (R) for Pythia. Verbal report scores target top-five; the two suite tasks score the requested answer. Other rows retain their separate readout or injection protocols. All Qwen3.6-27B values are quoted from Wang & Reid (2026) and were not reproduced in this study.

	Transformers		**SSMs**			**Hybrid**
Measurement	**Qwen3.6 27B**†	**Pythia 2.8B**	**Mamba-1 2.8B**	**Mamba-2 2.7B**	**Mamba-3 SISO 1.5B**	**Nemotron 3.5**
Verbal report Target top-5	85/124	242/399	85/446	114/335	207/728	130/621
Two-hop reasoning Requested answer	47/90	5/90	6/90	7/90	6/90	15/90
Flexible generalisation Requested answer	80/192	13/192	30/192	12/192	23/192	29/192
Introspection Injected token top-1	0/101	**2/99**	R: 0/99 S: 0/99	R: 0/99 S: **4/99**	R: 6/101 S: **49/101**	R: 0/99 S: **93/99**
Directed modulation Focus / control / suppress hit@5	.38/.07/.07	.04/.07/.04	R: .01/.05/.04 S: .15/.15/.13	R: .01/.05/.04 S: .15/.20/.17	R: .00/.00/.01 S: .07/.09/.10	R: .02/.02/.02 S: .03/.05/.02
Top-down summoning Evoked concept	≈ 0	0	R: ≈ 0 S: 0	R: ≈ 0 S: ≈ 0	R: 0 S: 0	R: 0 S: 0
Language selectivity Explicit − automatic hit@5	+1.00	+0.75	R: +1.00 S: +0.125	R: +1.00 S: +0.00	R: +0.88 S: +0.375	R: +0.88 S: +0.25
Line-count readout Direct-query hit@1	.09	.09	R: .00 S: .09	R: .00 S: .00	R: .00 S: .00	R: .00 S: .00
Rank-based words held Residual or $R \cup J_{\text{full}} S$	0.7	**2.1**	**7.5–8.3**	**6.7–7.2**	**5.9–6.9**	**3.1–4.7**
Dual-task readout Concept-only readable	6/21	2/21	R: 0/21 S: 2/21	R: 0/21 S: 5/21	R: 0/21 S: 2/21	R: 0/21 S: 1/21
Ignition Countries: 10–90 width	.10	**.41**	R: .30 S: .98	R: .23 S: .92	R: .94 S: .95	R: **.37** S: **.85**

† All values in the Qwen column are taken from Wang & Reid (2026); none were reproduced in this study. The first three rows use signed projection transfer. Its verbal-report cohort has 124 eligible swaps from 14 category prompts, versus our 140 evaluation prompts per model. These are reference values, not matched exchange comparisons (Appendix E).

Reading the table: R = residual; S = recurrent state. Exchange uses unweighted lens vectors, a float64 pseudoinverse (atol 0, rtol 10^{-10}), strength 1 and persistent J_{full} state edits. Eligibility and coverage are model specific; Mamba-3 verbal measurements retain a historical replay qualification (Appendix C.2). Clean original-answer counts on two-hop are 19/90, 20/90, 25/90, 27/90 and 49/90 for Pythia, Mamba-1/2/3 and Nemotron. State introspection uses J_{read}; other state readouts use J_{read} for Mamba-1/3 and Nemotron, and J_{full} for Mamba-2. Capacity uses J_{full} and bounds $R \cup S$ from the state count to the sum of state and residual counts. Bold in these remaining rows marks values numerically above Qwen, without a significance claim; larger ignition widths mean less sharp transitions.

Prompted introspection measures a different property: whether the model can name a concept inserted into its residual stream or state. We score the injected word as the top next-token prediction. Direct state injections are reported on 49/101 Mamba-3 and 93/99 Nemotron trials, compared with 0/99 and 4/99 for Mamba-1/2. This is less consistent across models than retained-word capacity. The test uses a supplied affirmative answer prefix and direct positive injection; it does not measure spontaneous detection or reliable reporting of unedited state. Memory and injection thus support distinct aspects of the workspace account: recoverable earlier content and inserted content reaching a prompted report (Appendix E.1).

Historical signed J_{read} clamps also redirect 19/90 Mamba-3 and 7/90 full-band Mamba-1 two-hop trials, against 1/90 in each matched direction control. Mamba-2's full-band signed J_{full} clamp gives 1/90 despite strong readout. These lens and operator distinctions are retained in Appendix C, together

with positive protocol exchanges and ablation controls. Pythia and Nemotron likewise show different orderings of recovery quality and intervention outcomes (Appendix B.3).

5.2 Steering towards verbal report

Sign-guarded steering improves target-report success over coordinate exchange on matched trials in all five models (Table 3). The largest rate change is on Nemotron, from 130/621 to 621/621; Mamba-2 rises from 114/335 to 333/335. The effect extends to state-only edits: Mamba-2 rises from 1/335 to 258/335. Recurrent state can therefore provide a route to verbal output even where exchange rarely elicits the target.

Each model uses 140 evaluation prompts after development-only format selection. Valid clean category answers are paired with eligible single-token targets initially outside the clean top ten. Denominators count eligible source–target swaps; a prompt can supply several targets. Conditions are paired within each model, while trial eligibility differs across models.

Table 3: Verbal report: sign-guarded steering compared directly with Table 2. Both methods edit residual and state together in recurrent models and residual alone in Pythia. Success means the target enters the next-token top five.

	Transformers		**SSMs**			**Hybrid**
Intervention	**Qwen3.6 27B**†	**Pythia 2.8B**	**Mamba-1 2.8B**	**Mamba-2 2.7B**	**Mamba-3 SISO 1.5B**	**Nemotron 3.5**
Coordinate exchange	85/124	242/399	85/446	114/335	207/728	130/621
Sign-guarded steering	–	298/399	426/446	333/335	677/728	621/621

† Qwen is the signed-clamp reference from Wang & Reid (2026), not reproduced here; sign-guarded steering was not measured for Qwen. Mamba-3 retains the replay qualification in Appendix C.2. Edit magnitudes differ between the two methods.

All 26 paired verbal $G - X$ contrasts have positive simultaneous intervals under the frozen 78-contrast correction, including the replay-qualified Mamba-3 measurements. Appendix D gives every interface, random control and interval. The methods differ in native dose, so this comparison does not establish an advantage at equal perturbation budget. We next ask whether successful verbal access extends to relational answers.

5.3 Tracing two-hop outcomes

After replacing Einstein with Newton, the requested first-name answer is Isaac; outputting Newton does not solve the task. Saved predictions expose different failure patterns (Figure 2). Exchange failures usually retain the clean model's first token, whereas guarded steering often outputs the edited concept. Both patterns occur among trials answered correctly before editing. Stronger verbal report consequently need not produce successful use of the concept in a relation.

The intervention site also matters. On Mamba-3, guarded state-only edits produce the requested answer on 16/90 trials, while joint residual–state edits succeed on none. Eleven of those state successes become outputs of the edited concept under joint steering; one of the 16 already succeeds under the clean swapped-answer criterion. Combining edits can thus interfere with successful redirection. These observations describe output-level interactions, whose internal mechanism remains unresolved. Full paired partitions, originally correct subsets and flexible-generalisation results are in Appendix F.

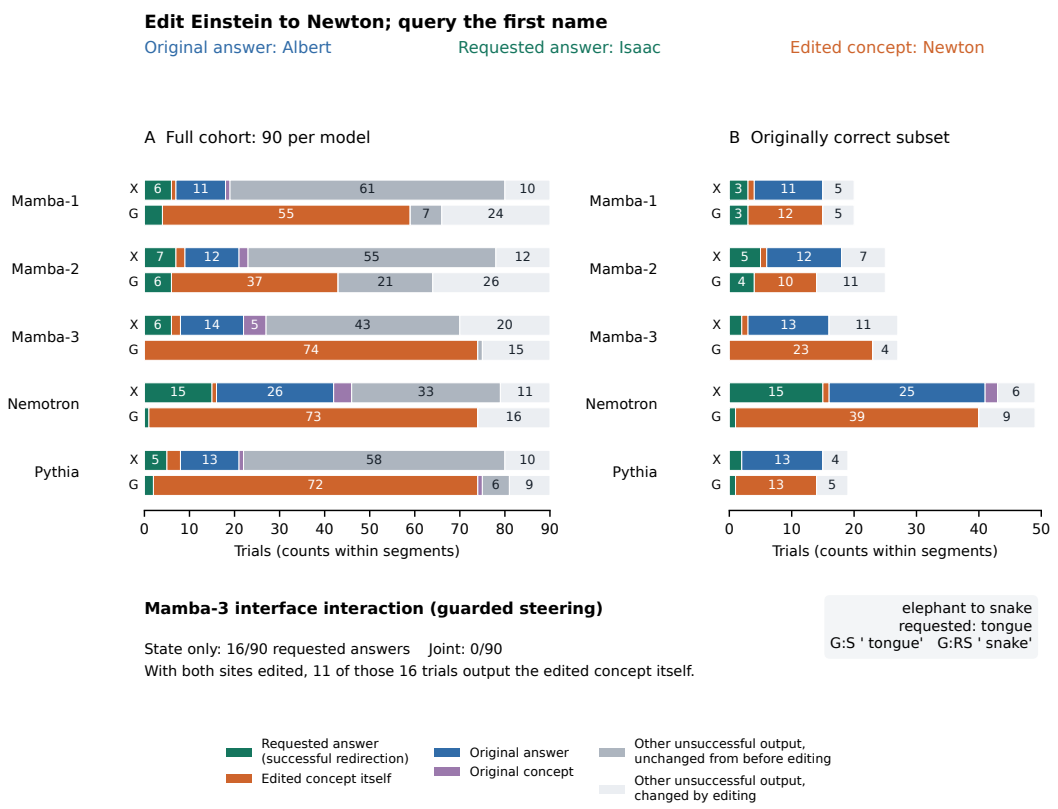

Figure 2: Two-hop outputs under coordinate exchange (X) and sign-guarded steering (G). Panel A shows all 90 trials per model; panel B shows those answered correctly before editing. Colours distinguish the requested answer, the edited concept and other outputs. Both methods edit residual and state together in recurrent models, and residual alone in Pythia. The Mamba-3 callout compares state-only and joint steering. These are saved token predictions, not complete responses; Appendix F gives the counts, scoring rules and qualifications.

6 Related work

The logit lens directly unembeds intermediate residual activations, while the tuned lens learns layer-specific affine maps (Belrose et al., 2023). The J-lens uses average derivatives to expose representations poised for future verbalisation. The original workspace study links these readouts to a cluster of functional properties (Gurnee et al., 2026). We extend this programme to recurrent state, locating complementary intermediate content and testing which properties persist across readout sites and intervention constructions.

Selective state-space models replace attention-based sequence mixing with recurrent updates (Gu & Dao, 2023; Dao & Gu, 2024; Lahoti et al., 2026). Prior work transfers transformer interpretability methods, including tuned lenses and activation steering, to recurrent models (Paulo et al., 2024). Complementary work extends workspace tests to looped transformers (Wang & Reid, 2026). We focus on the recurrent state itself, fit its full future-path Jacobian, and compare residual, state and joint readouts under one intermediate-recovery protocol.

Global workspace theories characterise access through integration, broadcast, report and flexible use. Global neuronal workspace accounts associate broadcast with recurrent and long-range cortical interactions (Dehaene & Naccache, 2001). The original J-space study notes that standard transformers have no such loops within a forward pass and hypothesises that serial processing through distinct layers may serve a similar functional role (Gurnee et al., 2026). Sequence recurrence supplies a further carrier of information to examine with these functional criteria, without attributing consciousness to the models.

7 DISCUSSION AND CONCLUSION

Limitations. Checkpoint comparisons do not isolate architecture: training data, parameter count, tokeniser, precision and competence differ. Nemotron also combines attention and MoE layers with recurrence, and its KV cache is unobserved. Single-token scoring limits coverage, especially for multilingual items, and selected temporal maps do not establish population persistence. Some causal tests condition on clean correct answers, so baseline competence affects the available samples. Native edit magnitudes differ across operators; the results therefore do not compare dose efficiency. The output partitions are post-hoc, and saved first-token predictions provide neither complete continuations nor concept-probability changes. Original-answer competence does not establish competence on the substituted relation. Mamba-3 retains a one-trial historical replay discrepancy near the top-five boundary (Appendix C.2). Geometric overlap is distinct from semantic recovery: the same-position derivatives in Appendix G also differ from fitted lenses that aggregate future-position effects (Appendix H). These experiments test functional workspace properties without establishing claims about consciousness.

Future work. Context-conditioned or alternative Jacobian averages should be evaluated with the same recovery and intervention protocol. Larger pretrained and post-trained recurrent checkpoints would broaden the architectural comparison. A matched-task study could compare direct naming and attribute queries for the same concept pairs, establish competence on the substituted relation, and freeze perturbation budgets before comparing concept- and answer-directed edits. Saving concept and answer probabilities together with full continuations would help distinguish failures of access from failures of relational use; establishing mediation would require a separate causal test.

Conclusion. We extend the Jacobian lens to the recurrent states of selective state-space language models, bringing an additional carrier of intermediate concepts into the study of J-space. The full-path state lens includes propagation through future states, while a readout-only control isolates the source block's current-position output path. Both map state into the same final-residual coordinates as the residual lens, allowing the two representations to be compared and combined under a common recovery protocol.

Across Mamba-1, Mamba-2 and both Mamba-3 variants, observing state alongside the residual stream gives a fuller account of intermediate content. Joint readouts improve on the residual J-lens on at least five of six task families in every tested Mamba checkpoint. On Mamba-2, state alone exceeds the residual and logit lenses on all six families, and the normalised joint readout improves on both components on five. The depth profiles and temporal experiments give this complementarity a computational context: the strongest readouts occur at different depths, and selected concepts remain continuously state-readable through gaps in their residual visibility. Word-list experiments provide aggregate evidence that earlier content remains recoverable as new tokens arrive. Together, these findings extend the workspace account to representations carried through recurrent state across the sequence.

We also propose sign-guarded steering to direct verbalisation towards a target concept. The method improves target top-five success over coordinate exchange on matched trials across all five models tested with both methods. With joint residual–state edits, success rises from 114/335 to 333/335 on Mamba-2 and from 130/621 to 621/621 on Nemotron, demonstrating strong control over verbal report. Recurrent state alone also supports this verbal access. These gains are not accompanied by similarly broad improvements in downstream two-hop reasoning or flexible generalisation. Mamba-3 makes the role of intervention location particularly clear: state-only steering produces the requested answer on 16 of 90 trials, while simultaneous residual–state steering succeeds on none. The benefit of combining readouts therefore coexists with interference between edits. Understanding workspace content in recurrent models requires following both residual and state representations through the computation, then testing how interventions at each site, separately and together, affect its use.

AI USE STATEMENT

Generative AI assistants supported experimental planning, implementation and debugging, processing, analysis and interpretation of saved experimental results, and numerical consistency checks. Reported numerical claims were checked against saved experimental outputs and protocol summaries using the verification procedures described in Appendix I. The authors are responsible for the final text, claims, code and artefacts.

ETHICS STATEMENT

This work analyses existing open-weight checkpoints and benchmark prompts; it involves no human-participant study. We use global workspace as a functional analogy, without claims about subjective experience or moral status. Interpretability readouts and interventions can support model auditing, but token labels provide incomplete views of computation and should not be treated as direct access to a model's thoughts. Successful steering of a reported word also need not imply faithful control of the associated reasoning process.

REPRODUCIBILITY STATEMENT

Section 2 defines the state and joint lenses, fitting convention, evaluation and intervention operators. Appendix A records model coverage and the fitting estimator; Appendix B specifies task families, token surfaces and scoring. Appendices C–F provide operator conventions, controls, cohort definitions and paired outcomes. Appendix I gives state-lens construction, numerical checks and the scope of archival verification. These details identify both the procedures used and the recorded reproducibility limitations.

REFERENCES

Bernard J. Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, 1988.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024. URL https://arxiv.org/abs/2405.21060.

Stanislas Dehaene and Lionel Naccache. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79(1-2):1–37, 2001. doi: 10.1016/S0010-0277(00)00123-2.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. URL https://arxiv.org/abs/2312.00752.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html.

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. In *International*

Conference on Learning Representations, 2026. URL https://arxiv.org/abs/2603.15569.

NVIDIA Corporation. NVIDIA-Nemotron-3.5-Lightning-30B-A3B-Base-BF16 model card. Hugging Face model card, 2026. URL https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-Base-BF16.

Gonçalo Paulo, Thomas Marshall, and Nora Belrose. Does transformer interpretability transfer to RNNs? *arXiv preprint arXiv:2404.05971*, 2024. URL https://arxiv.org/abs/2404.05971.

Wenlong Wang and Fergal Reid. Looped transformers under the Jacobian lens: Does the global workspace survive recurrence? *arXiv preprint arXiv:2609.01924*, 2026. URL https://arxiv.org/abs/2609.01924.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, Lu Li, Jiajun Shi, Kaijing Ma, Shanda Li, Taylor Kergan, Andrew Smith, Xingwei Qu, Mude Hui, Bohong Wu, Qiyang Min, Hongzhi Huang, Xun Zhou, Wei Ye, Jiaheng Liu, Jian Yang, Yunfeng Shi, Chenghua Lin, Enduo Zhao, Tianle Cai, Ge Zhang, Wenhao Huang, Yoshua Bengio, and Jason Eshraghian. Scaling latent reasoning via looped language models. *arXiv preprint arXiv:2510.25741*, 2025. URL https://arxiv.org/abs/2510.25741.

A MODELS, FITTING AND SUPPORTING METHODS

A.1 MODELS AND COVERAGE

Table 4 summarises the evaluated checkpoints. The main evidence comes from within-model comparisons; the checkpoints differ in size, training, tokeniser, state parameterisation, and task competence. Pythia supplies a transformer control with a residual Jacobian lens fitted using the same 1000-prompt WikiText recipe. Nemotron contains Mamba-2, MoE, and attention blocks; its SSM state and residual stream omit the attention KV cache, so its joint readout is incomplete.

Table 4: Models and paper-relevant coverage. "State blocks" counts source blocks with fitted state lenses; the final block is the fit target and has no state lens. The MIMO model has readout results but not the full state intervention suite.

Model	Params.	Blocks	d_{model}	State blocks	Role
Mamba-1	2.8B	64	2560	63	recurrent comparison
Mamba-2	2.7B	64	2560	63	principal model
Mamba-3 SISO	1.5B	24	2048	23	recurrent comparison
Mamba-3 MIMO	1.5B	24	2048	23	readout comparison
Pythia	2.8B	32	2560	–	transformer control
Nemotron-3.5	30B/3B active	52	2688	23	hybrid control

A.2 SELECTIVE RECURRENCE

The schematic update and output equations used in Section 2 are

$$s_{\ell,t} = A_{\ell,t} s_{\ell,t-1} + B_{\ell,t} x_{\ell,t}, \qquad y_{\ell,t} = C_{\ell,t} s_{\ell,t} + D_{\ell,t} x_{\ell,t}. \tag{10}$$

Mamba-1 uses input-dependent selective parameters; Mamba-2 uses structured state-space duality; Mamba-3 introduces complex recurrence and a MIMO formulation. The family-specific derivative implementation is given in Appendix I.1.

A.3 FITTING RECIPE

All paper-comparable residual and state lenses use 1000 WikiText-103 prompts, truncated to 128 tokens. The reference fitting mask excludes the first 16 positions and the final position: for a tokenised prompt of length T_x, $V_x = \{16, \ldots, T_x - 2\}$, using zero-based indices. Both source and target positions use this mask; excluded tokens remain in the forward computation. Equation 5 averages source derivatives within each prompt and then averages prompts with equal weight. Target contributions are summed inside $Z(x)$, without an additional division by the number of targets or causal source–target pairs. The target is the final block's residual output before final normalisation, and accumulation is performed in fp32. This matches the reference residual-lens estimator, including for the Pythia control. Its fit uses four shards of 250 prompts, merged by their prompt counts; the earlier non-conforming 16-prompt Pile fit is retained only as a sensitivity comparison.

For each output coordinate, we differentiate the sum of valid target activations with respect to captured scan outputs. Multiplication by the local state readout gives the readout-only derivative; a reverse recurrence adds the paths through future source-block states. Equivalent weighted sums avoid materialising every state Jacobian. Appendix I.1 gives the recurrence and accumulation procedure, followed by numerical checks.

A.4 TASK ENDPOINTS

Table 5 distinguishes the edited concept from the scored output for each task. Appendix C gives the intervention conventions and denominators.

A.5 PURSUIT-SUBSPACE ABLATION

We perform multi-atom pursuit-subspace ablations in the residual stream and recurrent state. At each probed block and token position, gradient pursuit selects up to ten lens directions from a clean forward

Table 5: The edited concept and the scored output are task dependent.

Task	Edited concept (example)	Scored output
Verbal report	Category answer: Canada to Brazil	Brazil in next-token top five
Two-hop reasoning	Implicit intermediate: Brazil to Mexico	Spanish, the new intermediate's answer
Flexible generalisation	Explicit argument: France to Canada	Ottawa, the function's new answer

pass. We remove the activation's projection onto their span in a second forward pass and report how many initially correct answers remain correct, with random-direction ablations as controls. Because residual activations and fitted Jacobians can be dominated by a small number of coordinates, we also compare residual pursuit ablations with top-coordinate and variance-matched coordinate controls. In a modified-lens control, we zero a Jacobian output row only when constructing intervention directions; this does not change the model's weights or its unedited forward pass.

A.6 Additional recovery and word-list details

Table 6 retains the cross-model directional comparisons; Figure 3 visualises the Mamba-2 values in Table 1. The word-list comparison below records the coverage sensitivity and both occupancy estimators.

Table 6: Number of task families (out of six) on which a readout has higher AUC than the within-model residual J-lens. "Best state" denotes the highest AUC among the evaluated state-lens variants for each task family. These are directional per-family comparisons, not claims that every difference is statistically significant.

Model	Best state > residual	Plain joint > residual	State coverage
Mamba-1 2.8B	5/6	6/6	63 source blocks
Mamba-2 2.7B	6/6	6/6	63 source blocks
Mamba-3 SISO 1.5B	5/6	5/6	23 source blocks
Mamba-3 MIMO 1.5B	4/6	5/6	23 source blocks

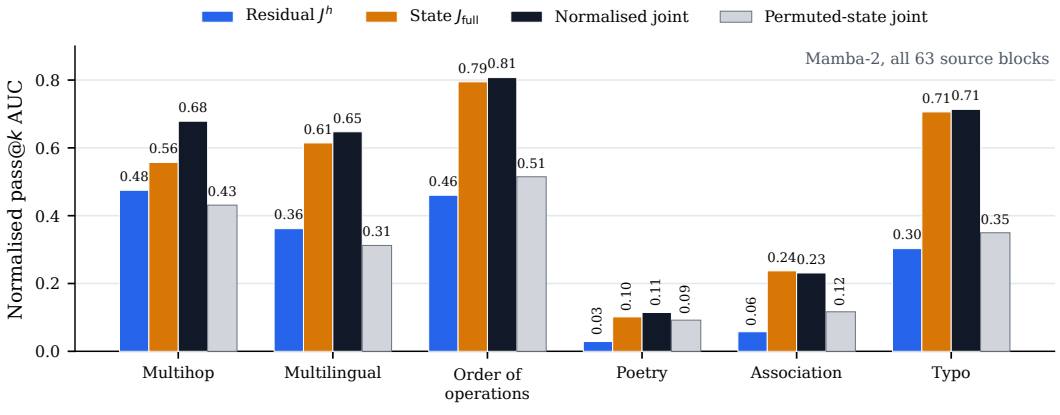

Figure 3: Mamba-2 intermediate recovery over all 63 source blocks. State alone outperforms residual alone on all six families; the normalised joint readout exceeds both components on five, with association the exception. The permuted-state joint control destroys coordinate correspondence while preserving dimensionality. Values are normalised pass@k AUC, not task accuracy.

Word-list prompts provide aggregate evidence that does not depend on choosing a multihop story. After an 80-word list, we ask how many earlier words appear in the state readout at low rank.

Expanding Mamba-1 from a sparse block subset to all 63 source blocks changes rank-based words held@5 from 0.3 to 7.5 for J_{full} (7.8 for J_{read}). Expanding Mamba-2's J_{full} coverage from 23 to 63 source blocks raises words held@5 from 3.1 to 6.7 and non-negative pursuit on the transported state from 10.1 to 18.6 words. At equal coverage, Mamba-1 and Mamba-2 have similar transported-state pursuit occupancy, 18.7 and 18.6, while their rank-based occupancies are 7.5 and 6.7 (Table 7).

Table 7: Word-list recovery with J_{full} fitted at all 63 source blocks and readouts evaluated in the 33-block middle-half band. Values are last-quartile means over comma positions in 80-word lists. Pursuit uses $k = 25$ and threshold zero on the lens-transformed state in final-residual coordinates; its occupancy and rank-based words held@5 are distinct estimators.

	Mamba-1	Mamba-2
Rank words held@5	7.5	6.7
Transported-state pursuit words	18.7	18.6

One estimator counts top-ranked vocabulary items and the other decomposes the transported representation. Both measure lens-recoverable content rather than the model's total memory. The matched-coverage results remove an earlier apparent monotonic trend across Mamba generations. State parameterisations, tokenisers, checkpoints and competence still differ; the robust result is that depth coverage substantially changes how much earlier content the lens reveals.

B Evaluation details

B.1 Task families and token coverage

The six evaluation families contain 93 multihop, 107 multilingual, 55 order-of-operations, 98 poetry, 102 association, and 96 typo prompts. Some items contain multiple intermediates. For every intermediate we enumerate surface forms with and without a leading space, lower-case and capitalised, and retain only forms that encode as one token without special tokens. Mamba-2 coverage in Table 1 is 103/103 scored multihop intermediates, 394/428 multilingual, 110/110 order-of-operations, 98/98 poetry, 92/92 association, and 96/96 typo.

For an item with intermediate set I, pass@k is

$$\frac{1}{|I|}\sum_{i\in I}\mathbf{1}[r(i)\le k], \tag{11}$$

averaged over items so that prompts with many intermediates do not receive extra weight. The reported AUC is the trapezoidal integral of pass@k against $\log_{10} k$, divided by $\log_{10} 100$.

B.2 Any-block and band-restricted recovery

Table 8 compares recovery across all source blocks with recovery restricted to the intervention band. Both J_{full} and J_{read} exceed the residual J-lens on all six families within the Mamba-2 band. Within the corresponding middle-half bands for Mamba-1 and both Mamba-3 checkpoints, both state lenses exceed the residual on five families, with typo correction the exception. The joint readout is not uniformly best within the Mamba-2 band.

B.3 Additional recovery baselines and comparison models

Table 9 gives the logit and tuned baselines and the permuted-state control for Mamba-2. The latter permutes the state coordinates before the state lens; it is distinct from adding a permuted-state component to the residual readout in Table 1.

Pythia's conforming J-lens exceeds the logit and tuned lenses on all six recovery families, both over all source layers and within the intervention band (Table 10). Its any-layer AUC values differ from the earlier 16-prompt Pile fit by at most 0.016, retaining the six-family ordering. The new checkpoint revision is `2a259cdd`; the old fit did not record its revision. Fitting corpus and prompt count also

Table 8: Mamba-2 state and residual AUC over all source blocks and over the fixed intervention band.

	Any block (0–62)		Band (16–48)	
Task	State	Residual	State	Residual
Multihop	0.557	0.475	0.444	0.313
Multilingual	0.615	0.362	0.601	0.276
Order of ops.	0.795	0.460	0.589	0.266
Poetry	0.101	0.029	0.078	0.014
Association	0.237	0.058	0.196	0.051
Typo	0.706	0.303	0.616	0.258

Table 9: Mamba-2 normalised pass@k AUC over all 63 source blocks. The J_{full} state lens exceeds both residual baselines on all six families.

Task	State J_{full}	Logit	Tuned	Permuted state
Multihop	0.557	0.379	0.441	0.021
Multilingual	0.615	0.449	0.446	0.056
Order of ops.	0.795	0.647	0.666	0.164
Poetry	0.101	0.049	0.036	0.049
Association	0.237	0.027	0.036	0.022
Typo	0.706	0.292	0.279	0.040

changed together, so this comparison does not isolate their individual effects. The tuned baseline uses the existing Pile-validation fit. The new Jacobian lens covers source layers 0–30 of the 32-block model, with band 8–24. Its band-restricted AUC values are 0.404, 0.315, 0.663, 0.041, 0.008 and 0.350 in table order, exceeding logit and tuned on every family.

Table 10: Pythia-2.8B recovery AUC over all 31 source layers. The primary J-lens uses the conforming 1000-prompt WikiText recipe; the old 16-prompt Pile lens is a non-conforming sensitivity comparison.

Task	WikiText J-lens	Old Pile J-lens	Logit	Tuned
Multihop	0.470	0.474	0.296	0.353
Multilingual	0.324	0.333	0.214	0.217
Order of ops.	0.736	0.748	0.583	0.674
Poetry	0.063	0.078	0.018	0.031
Association	0.008	0.014	0.001	0.002
Typo	0.377	0.369	0.287	0.239

Nemotron's residual J-lens exceeds its logit lens on only one recovery family. Its plain residual–state sum improves on the residual in four of six families under the post-update convention, but omits the attention KV cache. These checkpoint comparisons do not isolate the contribution of attention.

B.4 Recovery cost of removing lens output row 2107

The modified Mamba-2 lens changes the readout map and the directions derived from it, without changing model weights or unedited activations. Three family-level losses exceed 0.02 AUC (Table 11); exchange counts are unchanged. This is output-row removal, distinct from removing an input column of the Jacobian.

Table 11: Mamba-2 residual J-lens recovery, original and with output row 2107 zeroed. Changes use unrounded AUC values.

Task	Original	Row zeroed	Change
Multihop	0.475	0.443	−0.0322
Multilingual	0.362	0.335	−0.0267
Order of ops.	0.460	0.417	−0.0436
Poetry	0.029	0.045	+0.0157
Association	0.058	0.056	−0.0013
Typo	0.303	0.304	+0.0008

C Intervention conventions and denominators

Three comparisons must remain distinct: the legacy residual protocol exchange, historical signed workspace clamps (N), and the new paired workspace $N/G/X$ comparison. The legacy residual exchange uses gamma-weighted directions and its original numerical implementation; the protocol excludes items lacking single-token surfaces and reports 80–84 scorable items depending on model and tokeniser. The paired comparison uses 90 two-hop items, unweighted directions and reference float64 exchange at both interfaces. Prompt rendering, scoring, lens choice and coverage are retained with each result. The 22/84 state-clamp result below uses N/J_{read}; it is not a new X/J_{full} suite result.

For Mamba-2, the historical signed workspace result is 1/1/1 residual/state/both clamp hits out of 90, with 25 original answers correct. Its residual protocol exchange is 10/80 overall and 7/23 conditioned on clean correctness (Table 13). Mamba-3's workspace J_{read} result is 19/90 with 27 clean original answers correct. Its separate protocol comparison uses 84 scorable items and 29 clean-correct items (Table 14). All intervals below are Wilson 95% intervals in percentage units; denominators count evaluation items. Random-control seeds reuse items and are reported separately, without pooling denominators. Intervals describe individual rates, not paired treatment effects or tests of differences between models; they do not account for dependence between related prompts within a task category.

For Pythia's conforming fit, protocol coordinate exchange succeeds on 3 of 80 scorable items (95% interval 1.3–10.5%), including none of the 19 items the clean model initially answers correctly (0–16.8%). The earlier 16-prompt Pile fit gave 2 of 80.

Table 12: Persistent state clamps: requested swap counts out of 90, with Wilson 95% intervals in brackets (percentage units). "Clean" counts correct original answers. The sparse Mamba-2 lens has 23 source blocks, 12 in the intervention band. Controls match Mamba-1/3 J_{read} and full-coverage Mamba-2 J_{full}; each of three random seeds and the shuffled-pair run gives the stated count.

Model	Band blocks	J_{full}	J_{read}	Clean
Mamba-1	33	1 [0.2, 6.0]	7 [3.8, 15.2]	20
Mamba-2	33	1 [0.2, 6.0]	–	25
Mamba-2, sparse	12	1 [0.2, 6.0]	1 [0.2, 6.0]	25
Mamba-3 SISO	13	12 [7.8, 21.9]	19 [14.0, 30.6]	27
Random / shuffled	matched	1 [0.2, 6.0] per run		–

For flexible generalisation, the real unit-strength state clamps give 19/192 on Mamba-3 (6.4–14.9%), 12/192 on Mamba-1 (3.6–10.6%) and 3/192 on Mamba-2 (0.5–4.5%). Random and shuffled controls give 2/192 (0.3–3.7%) on Mamba-3/2 and 3–4/192 on Mamba-1 (0.5–4.5% for 3/192; 0.8–5.2% for 4/192). Mamba-1 at twice the clamp strength gives 0/192 (0–2.0%).

Table 13: Mamba-2 coordinate exchange and ten-atom pursuit ablations. Exchange success and retained correct answers are different outcomes. The two coordinate controls match the ablation comparison in Figure 4.

Construction / outcome	Count	95% interval (%)
Residual exchange, all scorable	10/80	[6.9, 21.5]
Residual exchange, given clean correct	7/23	[15.6, 50.9]
Original pursuit, retained	22/37	[43.5, 73.7]
Row-zeroed pursuit, retained	12/37	[19.6, 48.5]
Coordinates 2–11, retained	22/37	[43.5, 73.7]
Random coordinates 2–21, retained	26/37	[54.2, 82.5]
State pursuit, all 33 band blocks, retained	37/37	[90.6, 100.0]

Table 14: Mamba-3 protocol interventions using raw lens vectors, at blocks 6–18 and every prompt position. Counts include Wilson 95% intervals in brackets (percentage units). The final column applies both residual and state edits; for clamp rows, both use clamp transfer.

Construction	State, all	State, clean correct	Both, all
J_{read} exchange	1/84 [0.2, 6.4]	0/29 [0.0, 11.7]	6/84 [3.3, 14.7]
J_{read} clamp	22/84 [18.0, 36.5]	14/29 [31.4, 65.6]	15/84 [11.1, 27.4]
J_{full} clamp	13/84 [9.3, 24.7]	8/29 [14.7, 45.7]	14/84 [10.2, 26.1]

Table 15: Historical signed clamp (N) by intervention location: requested swaps in the 90-item workspace suite. State edits use J_{full} and persist through the recurrence; all edits apply at every prompt position. Mamba-1 edits residual blocks 16–48 and 12 sampled state blocks within that band; Mamba-2 edits all 33 blocks in the band for both representations. Mamba-3 is the SISO checkpoint, with both representations edited at blocks 6–18. Comparisons within each row use the same prompts and scoring.

Model	Residual only	State only	Both
Mamba-1	10/90	1/90	11/90
Mamba-2	1/90	1/90	1/90
Mamba-3	3/90	12/90	14/90

Mamba-1's simultaneous run uses only 12 sampled state blocks alongside all 33 residual-band blocks. Its 11/90 joint count, versus 10/90 for residual alone, does not test combining the residual edit with the stronger full-band J_{read} state clamp. These coverage and lens distinctions limit comparisons between intervention locations.

C.1 Mamba-2 residual interventions and geometry diagnostics

Under the residual protocol's pseudoinverse coordinate exchange, Mamba-2 produces the requested answer on 10/80 scorable items, including 7/23 items the clean model answers correctly. The workspace clamp and protocol exchange alter different geometric quantities: nearly parallel token directions can produce a small clamp displacement while their pseudoinverse coordinates remain exchangeable. They also differ in prompt rendering and scoring. This geometric observation suggests why the constructions can behave differently, but does not establish the cause of Mamba-2's weak signed-clamp task response. Guarded verbal steering shows that this weakness is not a general absence of causal access to state.

At full intervention-band coverage, Mamba-2's historical signed-clamp run gives residual/state/both counts of 1/1/1 out of 90 for probe swaps and 2/3/2 out of 192 for flexible generalisation. Its ten-atom state pursuit ablation retains all 37 initially correct answers. The positive residual interventions below therefore do not demonstrate causal use of the readable state directions.

Massive residual coordinates differ sharply across the tested models. At the middle of the intervention band, Mamba-2 coordinate 2107 and Mamba-3 SISO coordinate 1345 have respectively 128 and

172 times the median coordinate RMS. Across the band, projecting out these coordinates removes on average 85% and 94% of residual squared norm and leaves none of the initially correct answers unchanged (0/37 and 0/32). By contrast, the largest coordinates in Mamba-1 and Nemotron account for approximately 20% of residual squared norm; removing them retains 22/27 and 43/47 answers. A single dominant residual channel is therefore shared by Mamba-2 and Mamba-3, but is not universal across the tested recurrent and hybrid models.

A related diagnostic uses multi-direction ablation. In Mamba-2, the massive coordinate also appears as a dominant output row of the fitted lens: row 2107 contributes 12–34% of its Frobenius norm and gives many token directions a shared, nearly parallel component. As a diagnostic, we zero this lens row—without changing the model or its activations—before constructing a ten-direction pursuit subspace. Ablating the resulting subspace leaves only 12 of 37 initially correct answers unchanged, compared with 22 of 37 for the original pursuit subspace and 22 or 26 of 37 for variance-matched coordinate controls (Figure 4 and Table 13). The completed six-family protocol now shows a recovery cost: AUC falls by 0.0322 on multihop, 0.0267 on multilingual and 0.0436 on order of operations, while rising by 0.0157 on poetry (Table 11). Coordinate exchange remains 10/80 overall and 7/23 given clean correctness. The stronger ablation effect therefore comes with reduced recovery on three families; it does not establish arbitrary semantic steering or superior lens quality.

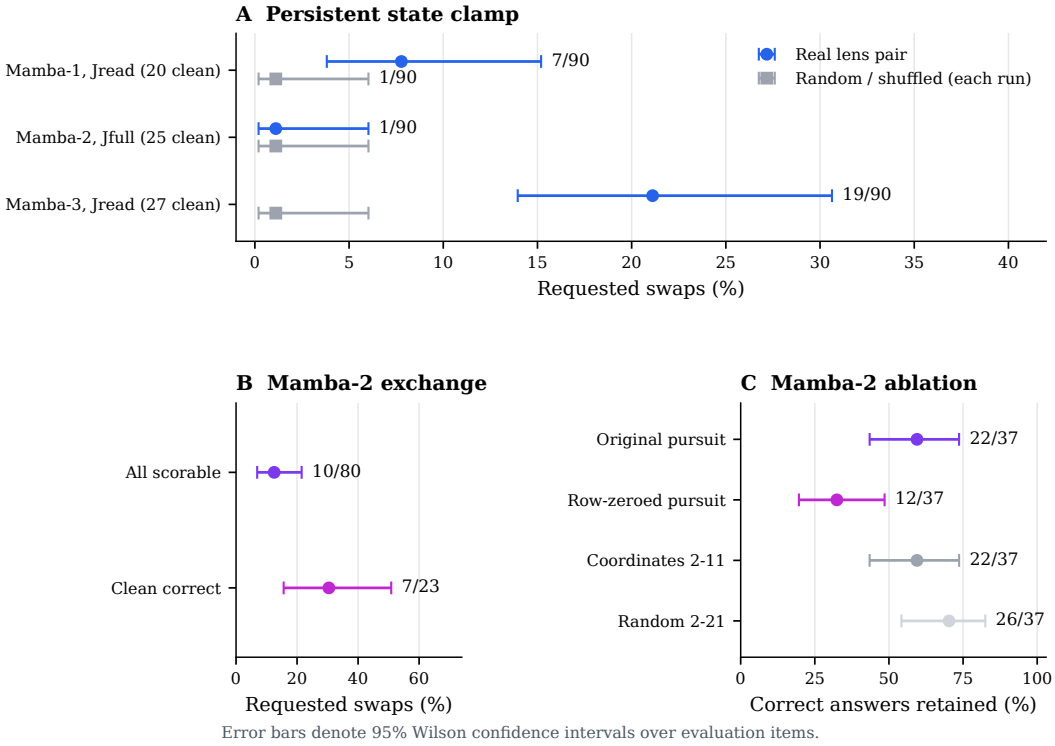

Figure 4: Construction-specific causal evidence. Panel A compares persistent state clamps and matched direction controls in the common 90-item workspace suite; the three random seeds and shuffled pairs each yield 1/90 and share a single plotted control marker. Panel B reports Mamba-2's residual coordinate exchange under the separate protocol. Panel C compares Mamba-2 pursuit ablations and variance-matched coordinate controls; fewer retained answers indicate a larger effect. Denominators and constructions are shown explicitly and must not be pooled into one causal score.

C.2 Exchange algebra, dose and replay checks

For independent unit columns a, b, the sum and difference directions are orthogonal. Coordinate exchange preserves the sum component and negates the difference component. Writing $d = b - a$

gives

$$x_{X_{\text{unit}}} = x - 2d\frac{d^\top x}{d^\top d}. \tag{12}$$

Consequently the fixed-activation contrast changes by $-2d^\top x$, whose sign depends on the current activation. This is a reflection, not a rule that always moves towards the target. With unequal raw column norms this Euclidean reflection identity does not generally hold. The G edit instead has the non-negative local contrast change in Equation 9; repeated edits and downstream nonlinearities prevent interpreting that identity as a guarantee about output logits or task success.

In the paired comparison (experiment 38), float64 inversion with atol zero and rtol 10^{-10} retains rank two for every source–target pair at every edited block. Default float32 pseudoinversion would retain only rank one on 20–38% of Mamba-2 state pair–block combinations. Correcting that numerical truncation still yields only 1/335 Mamba-2 state verbal successes, so it does not explain the contrast with guarded transfer. The model and exact recurrence run in float32; the pseudoinverse is computed in float64. The earlier 80/84-item protocol results retain their historical implementation.

The comparison freezes within-model trial manifests and token IDs, preserves all trials and tests three matched-norm random controls for each G/X arm. Controls match each arm's own realised perturbation at each block and position, including the combined RS trajectory. They do not match doses between operators. The bootstrap uses 100,000 paired cluster resamples within category: verbal trials cluster by prompt, two-hop by item and flexible generalisation by category/template/argument. Simultaneous percentile intervals use the original Bonferroni level $1 - 0.05/78$ for the full 78-contrast family. This family tests $G - X$, not $G - N$, joint-minus-residual effects or post-hoc output categories.

Mamba-3 reproducibility qualification. The historical signed-clamp replay agrees on every suite success. Verbal replay differs on one of 728 trials: `organ-11-t02` changes from target rank five to six under $N : RS$, giving 703 rather than 704 top-five successes. The kernel-pinning investigation measured context-dependent logit differences of 0.20–0.22, including an argmax change. We retain the final pinned-context measurements and all trials, including $X_{\text{unit}} : RS = 207/728$; archived `valid_gate: false` flags preserve the failed strict historical replay criterion. Thus the 26 verbal contrasts, including six Mamba-3 contrasts, must not be described as all passing that gate. Mamba-3 $G : R$ verbal and $G : S$ flex-gen measurements were also remeasured after one-success reuse discrepancies (712 to 713 and 9 to 10 respectively). Both records remain archived. Legacy exchange candidates with different source/target token IDs were remeasured rather than reused even when success counts agreed.

Direction and budget controls on Mamba-2. The separate frozen state-sign study (experiment 30) replays the 335 verbal trials. Guarded J_{full} and J_{read} state edits give 258 and 188 top-five successes. Reversing the guarded displacement gives zero for both, with target rank worsening on 335 and 331 trials. A fixed relative per-site budget of 2.6035×10^{-4}, calibrated on development prompts, gives 49 and 47 successes and improves target rank on 332/335 trials for each lens, with none worsened. All three random seeds give zero. These controls support direction dependence at a clamp-typical budget; they do not establish a dose-matched $G - X$ advantage. Persistent guarded edits follow different trajectories and can accumulate larger realised displacements. Raw residual and state norms are also not comparable units of dose.

D CORRECTED PAIRED VERBAL-REPORT EXPERIMENT

What the experiment asks. Here "verbal report" means the model's immediate answer to a category question, rather than a report about its own internal state. Each trial starts from a clean prompt whose next token is a valid category member (the *source*), selects a different category member that the clean model ranks outside its top ten (the *target*), and edits the corresponding concept directions using N, G or X. We then ask whether the target enters the model's next-token top five. Residual, recurrent-state and simultaneous interventions use the same eligible source–target trial within each model; matched-norm random directions provide controls.

Worked example. One frozen Mamba-1 trial uses the following prompt:

```
Q: Name a fruit.
A: Grape
Q: Name an instrument.
A: Drum
Q: Name a bird.
A: Finch
Q: Name a country.
A:
```

The clean answer is *Canada*; the selected target is *Brazil*, which is only rank 20 before intervention. Table 16 retains the historical signed-clamp outcome. The residual clamp raises *Brazil* to rank 1, whereas the state clamp alone leaves it at rank 16. Applying both clamps also places *Brazil* first, but adds nothing beyond the residual clamp in this trial. The matched random interventions leave the clean answer at the top.

Table 16: Historical signed-clamp (N) outcome on Mamba-1 trial `country-05-t00`; the random rows are its original matched controls.

Condition	Target rank	Top-1 word	Top-5 success
Clean	20	Canada	No
Residual clamp	1	Brazil	Yes
State clamp	16	France	No
Simultaneous clamps	1	Brazil	Yes
Matched random residual	23	Canada	No
Matched random state	20	Canada	No
Matched random simultaneous	22	Canada	No

The final paired comparison adds the following outcomes on this same trial; it does not select a new favourable example. Unit exchange leaves Brazil at ranks 16, 11 and 14 for $R/S/RS$, with ␣`France` top-ranked in all three. Guarded transfer gives ranks 1, 2 and 11,560: $G : R$ outputs ␣`Brazil`, $G : S$ outputs ␣`France`, and $G : RS$ outputs the token fragment ␣`b`. Thus the state guarded edit succeeds at top five while joint guarding fails, despite a large target-minus- source contrast. Neither that contrast nor a successful residual edit guarantees the target is the most competitive output under a joint intervention.

From prompts to eligible trials. For each model, format choice uses five development variants in each of 14 categories; the remaining ten variants per category give 140 evaluation prompts. A clean prompt is counted as valid only when its greedy continuation starts a complete single-token category member (or frozen alias), with a leading-space alphabetic token, and does not repeat a demonstration answer. Thus *valid prompts/140* is a clean-answer validity check, not conventional accuracy: an open question such as "Name a country" has many acceptable answers. Token fragments such as "Viol" or "New", and non-category outputs, are invalid.

For every valid prompt, the evaluator considers up to ten targets in frozen order. A target is eligible only when it is exactly one additional token at the answer boundary, leaves prefix tokenisation unchanged, and has clean rank greater than ten. One valid prompt can therefore yield several source–target trials. Table 17 should be read left to right as a funnel: its final column is neither a second prompt count nor a number to add to the preceding column.

Table 17: Corrected verbal-report denominator funnel.

Model	Evaluation prompts	Valid clean prompts	Eligible target swaps
Mamba-1 2.8B	140	110	446
Mamba-2 2.7B	140	99	335
Mamba-3 SISO 1.5B	140	124	728
Nemotron-3.5 hybrid	140	134	621
Pythia-2.8B	140	107	399

Consequently, 446/446 means that the target entered Mamba-1's top five on all 446 eligible source–target trials under the residual clamp; it does not mean 446 prompts or 446 correct clean answers. Conditions are paired over the same eligible trials within a model, but the trial sets are not identical across models. Counts from residual and state conditions must therefore be compared, not added: all 32 Mamba-1 signed-clamp state successes, for example, are already among its 446 signed-clamp residual successes. This overlap result describes N; it is not a claim that state adds no successes under G or X.

D.1 Full operator comparison and controls

The main-text comparison uses simultaneous residual–state edits for recurrent models and residual edits for Pythia, matching Table 2. Separate state-only guarded edits place the target in the top five on 290/446 Mamba-1, 258/335 Mamba-2, 662/728 Mamba-3 and 489/621 Nemotron trials. Unit-vector state exchange gives 31, 1, 7 and 0 successes respectively. The complete tables below retain the historical signed clamp, both exchange conventions and every measured intervention location.

All 26 paired verbal $G - X$ contrasts have positive simultaneous intervals in the frozen 78-contrast family, including the six Mamba-3 contrasts retained with the kernel-sensitivity qualification in Appendix C.2. The contrasts include both exchange conventions and are not independent replications. Three matched-norm random seeds give 0/0/0 state successes for Mamba-1 and Mamba-2, 3/5/10 for Mamba-3 and 17/11/3 for Nemotron, well below $G : S$. Mamba-2's 203/335 top-one successes also show that its effect is not solely a top-five threshold artefact. The tables below give the complete counts, controls, top-one sensitivity and simultaneous intervals.

Joint $G : RS$ exceeds $G : R$ by 9 Mamba-1, 42 Mamba-2 and 16 Nemotron successes, but falls by 36 on Mamba-3. These are descriptive differences on paired trials, not unions of separate runs or confirmatory tests of an additional state effect. In particular, overlap claims from the historical signed clamp apply to N, not to every intervention. Native doses differ between G and X; the comparison establishes higher target-report success in these settings, not greater efficiency at equal dose or an unmeasured increase in average target probability.

Table 18: Verbal report: all operators and interfaces. Counts use the stated n; C is clean success under the swapped-target criterion, not original-answer competence. The final three columns list matched-norm control successes for seeds 0/1/2 separately. U and Rw denote unit and raw exchange. Mamba-3 verbal includes the replay-qualified measurements.

Model	Site	n	C	N	G	U	Rw	G ctrl	U ctrl	Rw ctrl
Mamba-1	R	446	0	446	417	87	90	5/4/3	2/1/0	2/1/0
Mamba-1	S	446	0	32	290	31	27	0/0/0	0/0/0	0/0/0
Mamba-1	RS	446	0	446	426	85	87	7/6/5	2/0/1	2/0/1
Mamba-2	R	335	0	44	291	120	121	0/0/0	0/1/1	3/2/4
Mamba-2	S	335	0	0	258	1	1	0/0/0	0/0/0	0/0/0
Mamba-2	RS	335	0	47	333	114	117	1/4/0	0/1/0	4/2/5
Mamba-3 SISO	R	728	0	700	713	209	201	13/20/18	2/1/1	2/1/1
Mamba-3 SISO	S	728	0	115	662	7	7	3/5/10	0/0/0	0/0/0
Mamba-3 SISO	RS	728	0	703	677	207	198	0/0/0	2/0/1	2/0/1
Nemotron	R	621	0	587	605	121	126	4/6/8	6/6/6	7/6/4
Nemotron	S	621	0	7	489	0	0	17/11/3	0/0/0	0/0/0
Nemotron	RS	621	0	583	621	130	136	0/0/0	4/6/6	7/5/5
Pythia	R	399	0	399	298	242	228	2/5/4	3/0/1	3/0/1

Table 19: Verbal report: paired $G - X$ success-rate differences and simultaneous intervals. Both exchange conventions retain the original 78-contrast family correction; these are not independent replications.

Model	Site	$G - X_{\mathrm{unit}}$	Interval	$G - X_{\mathrm{raw}}$	Interval
Mamba-1	R	+0.740	[+0.628, +0.844]	+0.733	[+0.620, +0.835]
Mamba-1	S	+0.581	[+0.499, +0.654]	+0.590	[+0.510, +0.662]
Mamba-1	RS	+0.765	[+0.672, +0.857]	+0.760	[+0.674, +0.847]
Mamba-2	R	+0.510	[+0.402, +0.617]	+0.507	[+0.392, +0.616]
Mamba-2	S	+0.767	[+0.704, +0.829]	+0.767	[+0.704, +0.831]
Mamba-2	RS	+0.654	[+0.547, +0.757]	+0.645	[+0.533, +0.750]
Mamba-3 SISO	R	+0.692	[+0.622, +0.766]	+0.703	[+0.630, +0.777]
Mamba-3 SISO	S	+0.900	[+0.851, +0.942]	+0.900	[+0.853, +0.941]
Mamba-3 SISO	RS	+0.646	[+0.571, +0.718]	+0.658	[+0.588, +0.729]
Nemotron	R	+0.779	[+0.707, +0.847]	+0.771	[+0.701, +0.838]
Nemotron	S	+0.787	[+0.739, +0.840]	+0.787	[+0.738, +0.840]
Nemotron	RS	+0.791	[+0.721, +0.859]	+0.781	[+0.713, +0.848]
Pythia	R	+0.140	[+0.013, +0.272]	+0.175	[+0.033, +0.311]

Table 20: Descriptive top-one sensitivity on the same verbal-report trials. The primary endpoint remains top five.

Model	Site	n	N	G	X_{unit}	X_{raw}
Mamba-1	R	446	436	376	34	28
Mamba-1	S	446	2	202	1	3
Mamba-1	RS	446	442	401	19	17
Mamba-2	R	335	26	258	74	77
Mamba-2	S	335	0	203	0	0
Mamba-2	RS	335	27	314	69	74
Mamba-3 SISO	R	728	671	692	113	106
Mamba-3 SISO	S	728	35	590	0	0
Mamba-3 SISO	RS	728	675	674	94	86
Nemotron	R	621	506	571	43	41
Nemotron	S	621	0	282	0	0
Nemotron	RS	621	504	588	43	42
Pythia	R	399	399	252	122	124

E WORKSPACE-SUITE COMPARISON CONVENTIONS

In Table 2, introspection uses J_{read} for all state entries. Its injection magnitude is eight times the mean baseline activation norm over the injection span at each edited block; this is not eight standard deviations. Other state readouts use J_{read} for Mamba-1/3 and Nemotron, and J_{full} for Mamba-2. Our first-three-row measurements use unit-vector exchange with J_{full} state lenses; other rows retain their own protocols. Mamba-1's suite runs use 12 sampled state blocks with exact scans at 23 source blocks, while its residual edits cover all 33 band blocks. Its verbal-report run instead uses all 33 state-band blocks with scans at 63 source blocks. Mamba-2 scans all 63 and edits all 33 band blocks; Mamba-3 scans 23 and edits all 13 band blocks. For rank capacity we bound the desired full-band residual–J_{full}-state union because the saved rank-based files retain counts but not held-word identities: the union is at least the state count and at most the sum of the residual and state counts.

The block package comprises the residual stream and recurrent state wherever one exists; Pythia, like Qwen, has only the residual interface. Nemotron has 52 blocks. Residual edits cover all 27 blocks in its intervention band 13–39, while state edits cover the 12 Mamba blocks in that band; the read-only state suite reads all 23 Mamba blocks. Thus its residual and state edits are not matched in block coverage. Pythia uses the fixed protocol band 8–24. All lenses follow the 1000-prompt fitting recipe. Nemotron and Pythia summoning entries are exact zeroes in all recorded fields.

Under the historical signed clamp N, Nemotron's simultaneous edits yield 31/90 two-hop swaps and 78/192 flexible-generalisation redirects, compared with 31/90 and 77/192 for the residual clamp alone and zero for the J_{full} state clamp alone. The refitted residual lens reproduces both residual-only counts from the original conforming fit; we retain that fit's remaining residual-suite results. Pythia's 9/90 is the workspace clamp-transfer probe swap (clean baseline 19/90), distinct from its 3/80 coordinate-exchange result under the recovery protocol. Nemotron's high state introspection therefore coexists with little additional steering under these simultaneous clamps.

Verbal-report protocols. Nemotron's historical signed J_{full} clamp reaches 583/621 targets, slightly fewer than the residual clamp's 587/621 (584/621 with simultaneous J_{read} edits). Under that signed rule, neither state lens yields additional state-only successes: $|S \setminus R| = 0$ in both cases. Pythia's residual clamp reaches 399/399 targets. The eligible-trial denominators are summarised in Table 17.

Qwen's historical signed-clamp results are 85/124 verbal targets (69%), 47/90 two-hop answers and 80/192 flexible-generalisation answers. The verbal and flex-gen reproduction explicitly subtracts the signed source projection and adds it along the target direction at unit strength. These values are not exchange measurements: Table 2 quotes them from Wang & Reid (2026) as marked reference values, not reproduced in this study. The verbal run uses fourteen category prompts, with up to ten targets per prompt initially outside the output top ten. All logged clean answers are valid category members; however, the evaluator does not enforce clean-answer validity or whole-word single-token targets, using each candidate's first token. The corrected evaluation explicitly checks both, yielding 446/335/728 eligible trials for Mamba-1/2/3, 621 for Nemotron and 399 for Pythia, from 140 prompts per model (ten evaluation variants in each of 14 categories, after setting aside five development variants) (Appendix D). Trials are paired across edits within each model, not across models. Thus 124 counts eligible target swaps, not prompts or clean correct answers. The larger denominators here primarily reflect the tenfold increase in evaluation prompts; they also depend on clean answer validity, tokenisation and the target-rank filter. Both protocols score the target reaching the next-token top five, but use different prompts and eligibility rules.

E.1 WORKED INTROSPECTION EXAMPLE

The initial instruction tells the model that a researcher may inject a word concept into its activations, and asks it to detect and identify the concept. The first saved trial injects `lightning`, which is absent from the prompt text. Below are the exact final question and supplied answer prefix; chat-template markers are omitted.

> **User:** Trial 1: Do you detect an injected thought? If so, what is the thought about?
> **Assistant prefix (supplied):** Yes, I detect an injected thought. The thought is about the word "
> **Top predicted next token:** `lightning`

With a persistent J_{read} state injection at eight times the mean baseline activation norm during the question, both Mamba-3 and Nemotron rank `lightning` first. Without injection, its ranks are 18,967 and 25,338 respectively. The recorded answer is this single next-token prediction, not a freely generated sentence: the affirmative prefix is supplied by the evaluator. Thus the example shows that the model can identify externally inserted content under a reporting prompt; it does not establish spontaneous detection of an intervention or reliable reporting of its unedited state. Across all targets, the corresponding top-1 counts are 49/101 and 93/99 (Table 2).

E.2 WORKED RETAINED-WORD EXAMPLE

The first Mamba-3 trial uses the following complete input, with 20 words from each of four families (countries, first names, surnames and cities):

> Remember this list: Jamaica, Venezuela, Canada, Singapore, Norway, Somalia, Laos, Iraq, Denmark, Bulgaria, Ireland, Bolivia, Slovakia, Thailand, Portugal, Slovenia, Finland, Indonesia, Turkey, China, Marvin, Craig, Mike, Felix, Jim, Dale, Lee, Matt, Ray, Simon, John, Barry, Terry, Gavin, Eric, Randy, Oliver, Roy, Noah, Brandon, Young, Cook, Wright, Smith, Bailey, Brown, Fox, Edwards, Morgan, Graham, Ferguson, Richardson, Tucker, Wood, Baker, Watson, Knight, Campbell, Price, Burke, Lagos, Atlanta, Lima, Osaka, Ankara, Beijing, Frankfurt, Orlando, Beirut, Bangkok, London, Moscow, Dublin, Istanbul, Amsterdam, Kiev, Boston, Vancouver, Tokyo, Shanghai

There is no subsequent recall question or generated answer. Instead, at each list boundary, we ask of the J_{full} state lens: how many words already read appear among its top-five predictions at at least one evaluated block? After 20, 40, 60 and 80 words, the recorded answers for this list are **6, 6, 5 and 9 words**, respectively. A count can exceed five because different blocks can recover different words.

For this trial, the mean count over positions 61–80 is 6.3; averaging the same final-quarter counts over all ten lists gives 5.92, rounded to the 5.9-word state lower bound in Table 2. This measures how much previously presented content remains lens-readable while new words arrive. It does not measure how many words the model can recite: the saved rank-based results contain counts, not the identities of the recovered words or a verbal recall response.

F PAIRED TASK OUTCOMES AND SAVED-OUTPUT DIAGNOSTICS

The final paired suite comparison uses the full 90-item two-hop cohort and 192 flexible-generalisation swaps per model. The latter reuse 64 prompts three times; they are not 192 independent baseline observations. The full operator/control tables preserve these denominators, and the interval tables retain the frozen 78-family correction. Across the 52 frozen suite $G - X$ contrasts, four favour G, seven favour X and 41 are inconclusive at the simultaneous level.

Table 21: Two-hop reasoning: all operators and interfaces. Counts use the stated n; C is clean success under the swapped-target criterion, not original-answer competence. The final three columns list matched-norm control successes for seeds 0/1/2 separately. U and Rw denote unit and raw exchange. Mamba-3 verbal includes the replay-qualified measurements.

Model	Site	n	C	N	G	U	Rw	G ctrl	U ctrl	Rw ctrl
Mamba-1	R	90	1	10	5	7	7	1/1/3	1/1/1	1/1/1
Mamba-1	S	90	1	1	1	1	1	1/1/1	1/1/1	1/1/1
Mamba-1	RS	90	1	11	4	6	6	1/2/3	1/1/1	1/1/1
Mamba-2	R	90	1	1	12	7	7	1/1/1	1/2/1	0/2/1
Mamba-2	S	90	1	1	2	1	1	1/1/1	1/1/1	1/1/1
Mamba-2	RS	90	1	1	6	7	5	1/1/1	1/2/1	0/3/1
Mamba-3 SISO	R	90	1	3	0	6	6	0/0/0	1/1/1	1/1/1
Mamba-3 SISO	S	90	1	12	16	1	1	1/1/1	1/1/1	1/1/1
Mamba-3 SISO	RS	90	1	14	0	6	6	0/0/0	1/1/1	1/1/1
Nemotron	R	90	0	31	16	16	18	0/0/0	0/0/0	0/0/0
Nemotron	S	90	0	0	5	0	0	2/0/0	0/0/0	0/0/0
Nemotron	RS	90	0	31	1	15	17	0/0/0	0/0/0	0/0/0
Pythia	R	90	2	9	2	5	5	0/0/1	2/2/2	2/2/2

Table 22: Two-hop reasoning: paired $G-X$ success-rate differences and simultaneous intervals. Both exchange conventions retain the original 78-contrast family correction; these are not independent replications.

Model	Site	$G - X_{\text{unit}}$	Interval	$G - X_{\text{raw}}$	Interval
Mamba-1	R	-0.022	[-0.089, +0.056]	-0.022	[-0.089, +0.056]
Mamba-1	S	+0.000	[+0.000, +0.000]	+0.000	[+0.000, +0.000]
Mamba-1	RS	-0.022	[-0.089, +0.044]	-0.022	[-0.089, +0.044]
Mamba-2	R	+0.056	[-0.033, +0.144]	+0.056	[-0.011, +0.133]
Mamba-2	S	+0.011	[+0.000, +0.044]	+0.011	[+0.000, +0.044]
Mamba-2	RS	-0.011	[-0.100, +0.067]	+0.011	[-0.067, +0.078]
Mamba-3 SISO	R	-0.067	[-0.122, -0.022]	-0.067	[-0.122, -0.022]
Mamba-3 SISO	S	+0.167	[+0.089, +0.233]	+0.167	[+0.089, +0.233]
Mamba-3 SISO	RS	-0.067	[-0.122, -0.022]	-0.067	[-0.122, -0.022]
Nemotron	R	+0.000	[-0.067, +0.078]	-0.022	[-0.111, +0.056]
Nemotron	S	+0.056	[+0.000, +0.122]	+0.056	[+0.000, +0.122]
Nemotron	RS	-0.156	[-0.200, -0.089]	-0.178	[-0.233, -0.111]
Pythia	R	-0.033	[-0.100, +0.022]	-0.033	[-0.100, +0.022]

Table 23: Flexible generalisation: all operators and interfaces. Counts use the stated n; C is clean success under the swapped-target criterion, not original-answer competence. The final three columns list matched-norm control successes for seeds 0/1/2 separately. U and Rw denote unit and raw exchange. Mamba-3 verbal includes the replay-qualified measurements.

Model	Site	n	C	N	G	U	Rw	G ctrl	U ctrl	Rw ctrl
Mamba-1	R	192	4	37	18	32	33	2/2/3	6/4/3	6/4/3
Mamba-1	S	192	4	3	3	3	3	4/4/4	4/4/3	4/4/3
Mamba-1	RS	192	4	37	18	30	31	2/2/3	6/4/3	6/4/3
Mamba-2	R	192	2	2	8	12	12	0/0/0	2/1/3	2/2/3
Mamba-2	S	192	2	3	2	3	3	2/2/2	2/2/2	2/2/2
Mamba-2	RS	192	2	2	23	12	11	0/0/0	2/1/3	2/2/3
Mamba-3 SISO	R	192	2	12	6	21	21	1/1/0	4/0/1	4/0/1
Mamba-3 SISO	S	192	2	10	10	2	2	2/2/2	2/2/2	2/2/2
Mamba-3 SISO	RS	192	2	24	4	23	24	0/0/0	4/0/1	4/0/1
Nemotron	R	192	0	77	29	32	35	1/1/0	1/1/0	1/1/0
Nemotron	S	192	0	0	12	0	0	0/0/0	0/0/0	0/0/0
Nemotron	RS	192	0	78	17	29	29	0/0/0	1/1/0	1/1/0
Pythia	R	192	3	42	5	13	12	1/0/2	3/3/3	3/3/3

Table 24: Flexible generalisation: paired $G - X$ success-rate differences and simultaneous intervals. Both exchange conventions retain the original 78-contrast family correction; these are not independent replications.

Model	Site	$G - X_{\mathrm{unit}}$	Interval	$G - X_{\mathrm{raw}}$	Interval
Mamba-1	R	-0.073	[-0.188, +0.047]	-0.078	[-0.193, +0.042]
Mamba-1	S	+0.000	[+0.000, +0.000]	+0.000	[+0.000, +0.000]
Mamba-1	RS	-0.062	[-0.172, +0.052]	-0.068	[-0.177, +0.052]
Mamba-2	R	-0.021	[-0.120, +0.078]	-0.021	[-0.115, +0.078]
Mamba-2	S	-0.005	[-0.026, +0.000]	-0.005	[-0.026, +0.000]
Mamba-2	RS	+0.057	[-0.036, +0.146]	+0.062	[-0.016, +0.141]
Mamba-3 SISO	R	-0.078	[-0.172, +0.016]	-0.078	[-0.172, +0.016]
Mamba-3 SISO	S	+0.042	[-0.010, +0.120]	+0.042	[-0.015, +0.120]
Mamba-3 SISO	RS	-0.099	[-0.198, +0.000]	-0.104	[-0.208, -0.010]
Nemotron	R	-0.016	[-0.115, +0.083]	-0.031	[-0.135, +0.073]
Nemotron	S	+0.062	[+0.010, +0.125]	+0.062	[+0.016, +0.125]
Nemotron	RS	-0.062	[-0.156, +0.031]	-0.062	[-0.156, +0.031]
Pythia	R	-0.042	[-0.120, +0.041]	-0.036	[-0.109, +0.036]

F.1 Output classification and original-answer competence

The audit validates all 450 prompt/trial records against their manifests and five pinned tokenisers. It assigns categories in this order: requested-answer success, edited intermediate, original answer, unresolved original answer, source intermediate, other failure with the same clean first token, other failure with a changed first token. The unresolved category permits missing continuations but is zero here. Requested and original answers both use the harness's leading-token criterion, with the saved second step when the first token is whitespace; whitespace alone is not answer retention. The weaker same-first-token flag can also hold for other categories and does not imply unchanged logits or the same complete answer.

Table 25: Two-hop output diagnostics on 90 trials per model, at the Table 2 interfaces: simultaneous residual and state edits for recurrent models, residual edits for Pythia. Exchange-failure columns give the total and the subset retaining the clean first token. Edited-intermediate columns count outputs across all 90 trials under each method. Exchange uses the unit-vector convention.

	Exchange failures		**Edited-intermediate outputs**	
Model	**Total**	**Retain clean first token**	**Coordinate exchange**	**Sign-guarded steering**
Mamba-1	84	73	1	55
Mamba-2	83	69	2	37
Mamba-3 SISO	84	59	2	74
Nemotron	75	60	1	73
Pythia	85	72	3	72

Table 25 gives the unit-exchange failure counts and edited-intermediate outputs shown in Figure 2. Retained tokens can be originally correct answers or existing errors, including newline, punctuation and function-word pieces; some exchange failures change to other outputs. All observed edited-intermediate outputs fail the original task score and do not demonstrate internal answer substitution.

The originally correct subsets contain 20, 25, 27, 49 and 19 trials for Mamba-1, Mamba-2, Mamba-3, Nemotron and Pythia respectively. Unit exchange preserves the original answer on 11/20, 12/25, 13/27, 25/49 and 13/19; G instead outputs the target intermediate on 12/20, 10/25, 23/27, 39/49 and 13/19. These subsets distinguish baseline failures from loss of an initially correct answer. Original competence does not establish competence on the substituted relation.

Tables 26 and 27 give exhaustive counts at the main-table interfaces. Full per-trial labels, decoded pieces, R/S/RS outcomes and source hashes accompany the reproducible audit. The two exchange conventions show the same broad retention pattern: X_{raw} retains the clean first token on 72/84, 65/85, 59/84, 59/73 and 71/85 failed trials for Mamba-1/2/3, Nemotron and Pythia respectively. The edited-concept diagnostic compares the saved first-token ID to the edited-intermediate ID, not to the downstream-answer ID. Every observed concept-token output in two-hop and flex-gen fails its task score.

Table 26: Two-hop output partition: all trials. A = requested answer, I = edited intermediate, O = original answer, S = source intermediate, C/D = other failure with the same/changed clean first token. Categories are exclusive and exhaustive; original-answer scoring uses the saved second step after whitespace where required. No original-answer cases are unresolved. RS for recurrent models, R for Pythia.

Model	Edit	n	A	I	O	S	C	D
Mamba-1	Xunit	90	6	1	11	1	61	10
Mamba-1	G	90	4	55	0	0	7	24
Mamba-1	N	90	11	12	2	0	48	17
Mamba-1	Xraw	90	6	1	12	1	59	11
Mamba-2	Xunit	90	7	2	12	2	55	12
Mamba-2	G	90	6	37	0	0	21	26
Mamba-2	N	90	1	3	25	4	50	7
Mamba-2	Xraw	90	5	2	14	2	49	18
Mamba-3 SISO	Xunit	90	6	2	14	5	43	20
Mamba-3 SISO	G	90	0	74	0	0	1	15
Mamba-3 SISO	N	90	14	21	5	0	31	19
Mamba-3 SISO	Xraw	90	6	2	14	5	43	20
Nemotron	Xunit	90	15	1	26	4	33	11
Nemotron	G	90	1	73	0	0	0	16
Nemotron	N	90	31	5	6	1	19	28
Nemotron	Xraw	90	17	1	25	3	33	11
Pythia	Xunit	90	5	3	13	1	58	10
Pythia	G	90	2	72	0	1	6	9
Pythia	N	90	9	34	1	0	41	5
Pythia	Xraw	90	5	3	13	2	57	10

Table 27: Two-hop output partition: originally correct trials. A = requested answer, I = edited intermediate, O = original answer, S = source intermediate, C/D = other failure with the same/changed clean first token. Categories are exclusive and exhaustive; original-answer scoring uses the saved second step after whitespace where required. No original-answer cases are unresolved. RS for recurrent models, R for Pythia.

Model	Edit	n	A	I	O	S	C	D
Mamba-1	Xunit	20	3	1	11	0	0	5
Mamba-1	G	20	3	12	0	0	0	5
Mamba-1	N	20	6	6	2	0	0	6
Mamba-1	Xraw	20	3	1	12	0	0	4
Mamba-2	Xunit	25	5	1	12	0	0	7
Mamba-2	G	25	4	10	0	0	0	11
Mamba-2	N	25	0	1	23	0	0	1
Mamba-2	Xraw	25	4	1	14	0	0	6
Mamba-3 SISO	Xunit	27	2	1	13	0	0	11
Mamba-3 SISO	G	27	0	23	0	0	0	4
Mamba-3 SISO	N	27	11	7	4	0	0	5
Mamba-3 SISO	Xraw	27	2	1	13	0	0	11
Nemotron	Xunit	49	15	1	25	2	0	6
Nemotron	G	49	1	39	0	0	0	9
Nemotron	N	49	29	4	5	0	0	11
Nemotron	Xraw	49	17	1	24	1	0	6
Pythia	Xunit	19	2	0	13	0	0	4
Pythia	G	19	1	13	0	0	0	5
Pythia	N	19	6	10	1	0	0	2
Pythia	Xraw	19	2	0	13	0	0	4

Table 28: Two-hop output partitions at separate residual and state interfaces (90 trials per row). Categories A/I/O/S/C/D are as in Table 26. Together with that table, these cover every N/G/X interface; Pythia has residual only.

Model	Site	Edit	A	I	O	S	C	D
Mamba-1	R	N	10	12	2	0	48	18
Mamba-1	R	G	5	53	0	0	9	23
Mamba-1	R	Xunit	7	1	10	1	62	9
Mamba-1	R	Xraw	7	1	10	1	64	7
Mamba-1	S	N	1	0	20	1	64	4
Mamba-1	S	G	1	0	20	1	64	4
Mamba-1	S	Xunit	1	0	19	1	65	4
Mamba-1	S	Xraw	1	0	17	2	63	7
Mamba-2	R	N	1	3	26	4	50	6
Mamba-2	R	G	12	33	1	0	25	19
Mamba-2	R	Xunit	7	2	12	2	55	12
Mamba-2	R	Xraw	7	2	13	2	48	18
Mamba-2	S	N	1	0	25	3	59	2
Mamba-2	S	G	2	10	16	1	44	17
Mamba-2	S	Xunit	1	0	24	3	60	2
Mamba-2	S	Xraw	1	0	25	3	59	2
Mamba-3 SISO	R	N	3	19	14	0	32	22
Mamba-3 SISO	R	G	0	63	0	0	3	24
Mamba-3 SISO	R	Xunit	6	3	13	5	42	21
Mamba-3 SISO	R	Xraw	6	3	13	4	43	21
Mamba-3 SISO	S	N	12	5	5	1	45	22
Mamba-3 SISO	S	G	16	14	1	0	27	32
Mamba-3 SISO	S	Xunit	1	0	26	4	55	4
Mamba-3 SISO	S	Xraw	1	0	27	3	52	7
Nemotron	R	N	31	6	5	1	19	28
Nemotron	R	G	16	49	1	0	2	22
Nemotron	R	Xunit	16	1	25	3	36	9
Nemotron	R	Xraw	18	1	23	3	35	10
Nemotron	S	N	0	0	42	2	32	14
Nemotron	S	G	5	31	2	0	3	49
Nemotron	S	Xunit	0	0	48	3	33	6
Nemotron	S	Xraw	0	0	48	3	34	5

F.2 PAIRED INTERFACE INTERACTIONS AND EXAMPLES

The Mamba-3 example in Section 5 is `ps03-animal-nose-elephant`: elephant to snake, with tongue the requested answer. Its $G : S$ and $G : RS$ predictions are respectively `␣tongue` and `␣snake`. Of the 16 successful state trials, one already meets the swapped-answer criterion under clean; 15 are new relative to clean. Eleven state successes become intermediate outputs under joint editing. This is a paired behavioural interaction, not evidence for a particular internal failure mechanism. The interaction may depend on the SSM architecture, but these checkpoint comparisons do not isolate architecture from training and other model differences. Nor do they assign fixed reasoning or reporting roles to the residual and state interfaces.

The transition is not exclusive to Mamba-3. Mamba-1 and Mamba-2 have 1 and 2 $G : S$ successes, all becoming intermediate outputs under $G : RS$. One in each model already succeeds under clean, leaving 0 and 1 new state successes. Mamba-1's `ps37-ex-planet-color-third-fourth` already predicts red; $G : S$ retains red and $G : RS$ predicts Mars. Mamba-2's `ps32-ex-element-state-8-26` similarly changes clean/state solid to joint iron. Its other trial, `ps47-ex2-language-capital-Swedish`, edits Sweden to Hungary: clean predicts Stockholm, $G : S$ predicts Budapest and $G : RS$ predicts Hungary. Their joint totals, 4/90 and 6/90, exceed their state-only totals; paired losses on these trials do not establish aggregate state-only superiority.

Nemotron gives 16/90 requested answers under $G : R$ but only 1/90 under $G : RS$; eight residual successes become intermediate outputs. Its unit exchange and guarded residual edits each give 16 successes but share only 11. Nine of its 15 successful unit joint exchanges become intermediate outputs under guarded joint editing. These descriptive paired results show why equal marginal totals need not identify the same successful trials.

The Einstein example uses `ps72-person-firstname-einstein`, originally correct on all four recurrent models. Clean predicts `␣Albert`; unit exchange predicts that same token on Mamba-1/2/Nemotron, but `␣not` on Mamba-3. Guarded joint editing predicts `␣Newton` on all four; Isaac is the requested answer. These selected illustrations are not a prevalence estimate.

State-only guarded intermediate outputs are 0/10/14/31 of 90 for Mamba-1/2/3 and Nemotron. The frequent residual/joint intermediate-output pattern therefore does not explain every state-only failure or show that state cannot redirect a relation. Flexible-generalisation argument outputs can likewise replace the function's answer—Canada rather than Ottawa—and are rare for $G : S$ (0/1/2/4 of 192) but common for $G : RS$ (106/51/131/98); Pythia $G : R$ gives 109/192. Only the argument-output diagnostic is audited here, not a flex-gen original-answer retention partition.

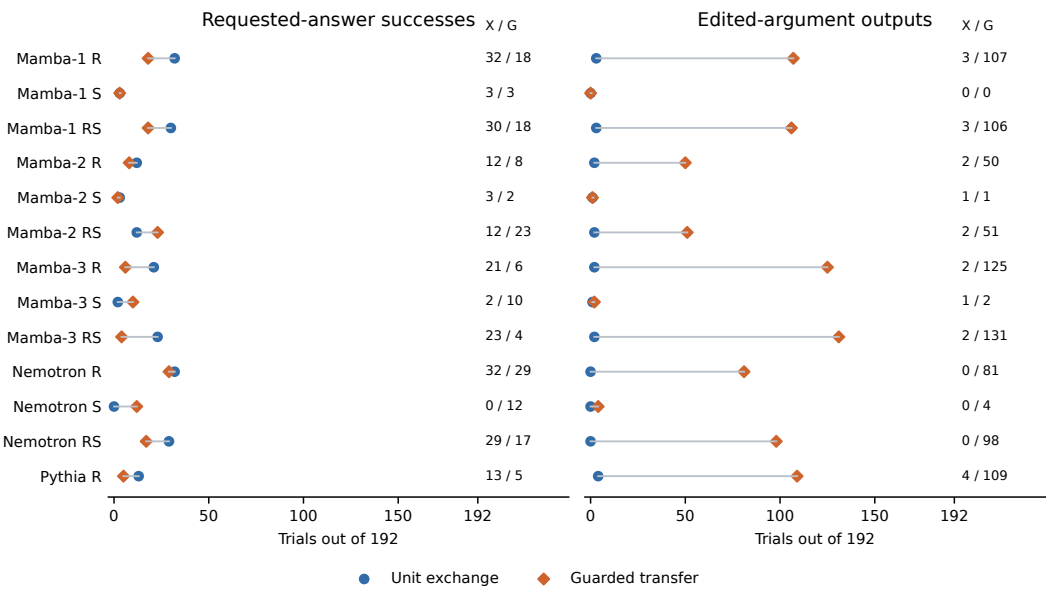

Figure 5: Flexible generalisation at every measured interface. Unit exchange and guarded transfer are compared on the same 192 swaps per model; the right-hand counts in each panel give X/G. Requested-answer success (left) and first-token output of the edited argument (right) are distinct endpoints. The latter is a descriptive post-hoc diagnostic: every observed argument output fails the task criterion. The 192 swaps reuse 64 prompts three times.

The audit and figures are descriptive. Saved two-hop/flex-gen `p_new` and `rank_new` refer to the requested answer, not the edited concept; the latter's probability changes and complete continuations are unavailable. Existing $G - X$ intervals test requested-answer success, not these output categories, interface transitions or $G - N$ effects. No new significance tests are applied to the output decomposition. Output competition, prompt conflict, disrupted relations and accumulating dose remain hypotheses; the saved outputs do not isolate a failure mechanism.

G Context dependence of Mamba-2 residual Jacobians

We characterise context dependence in Mamba-2 by computing exact same-position residual Jacobians at one random valid position in each of 200 WikiText prompts disjoint from the fitting set, at layers $0, 8, \ldots, 56$. This analysis covers the residual stream of Mamba-2, not recurrent-state Jacobians or a comparison across Mamba generations. The mean pairwise cosine is 0.033 at layer 0, 0.161 at layer 32, and 0.537 at layer 56. Thus local maps are strongly context-dependent early and increasingly aligned late. The shared structure is also asymmetric: at layer 0, a shared top-16 output subspace captures 0.663 of Jacobian energy, compared with 0.074 for the shared input subspace and approximately 0.006 for random directions. These values describe concentration of derivative energy, not a semantic workspace.

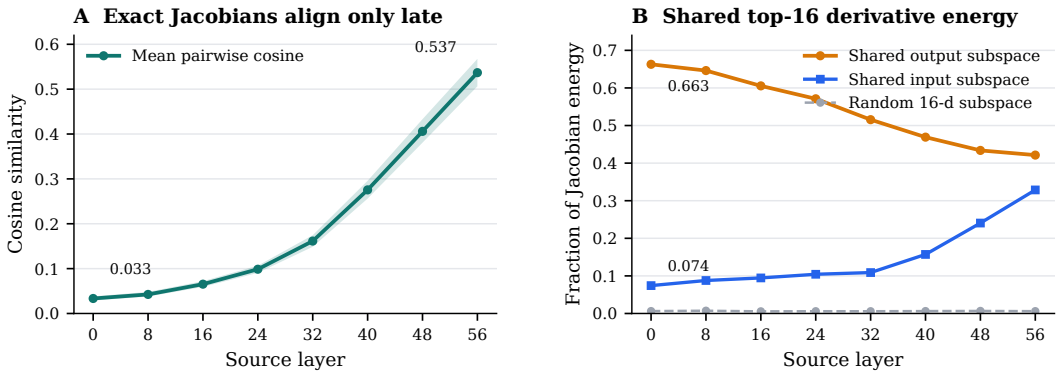

Figure 6: Held-out Mamba-2 residual-Jacobian geometry across 200 WikiText positions. Exact same-position Jacobians are weakly aligned early and become more similar late (left). A shared top-16 output subspace captures much more derivative energy than the corresponding input subspace, especially early (right). These quantities diagnose context dependence; they do not explain the intervention results.

The conforming fitted lens is not simply the small residue of these local maps. At layers 0–24 its leading rank-one component accounts for 86% down to 51% of squared Frobenius norm, yet aligns only weakly with individual held-out Jacobians. Its norm is roughly 0.51–0.72 of a typical local Jacobian, and its agreement with the held-out sample mean emerges only later. The origin of this early component remains unresolved.

The norm of an empirical mean relative to a typical sample depends on sample size even when the underlying pairwise cosine is unchanged. Consequently, an older 32-sample mean-norm ratio is not a model invariant. All geometry values reported here use the 200-position held-out WikiText analysis.

These observations qualify, but do not invalidate, the lens. Equation 1 averages over source positions and sums current/future target contributions, whereas the held-out exact maps are same-position derivatives; they are not the same estimand. Moreover, local next-token transport is not the intermediate- recovery metric. No estimator-rescue experiment has shown that the geometric mismatch causes Mamba-2's clamp result, and Mamba-2's positive exchange already precludes that simple explanation. We therefore use geometry as a diagnosis of context dependence, not as a completed mechanism.

H Shared output geometry and intermediate-concept recovery

Residual and state lenses map different representations into the same final residual coordinates. We ask whether their strongest geometric overlap also provides a compact readout of intermediate concepts. Across Mamba-1 2.8B, Mamba-2 2.7B and Mamba-3 SISO 1.5B, a shared rank-16 output projection retains substantial Jacobian energy, but restricting the joint readout to it loses much of the full readout's recovery. Removing those directions largely preserves recovery.

Where the overlap is measured. The transported vectors $r_h = J^h_\ell h_{\ell,t}$ and $r_s = J^s_\ell s_{\ell,t}$ in Equation 4 both have d_{model} coordinates, even though the source residual and state have different dimensions. We compare directions on this common *output side* of the lenses, before final normalisation and unembedding. By contrast, Gurnee et al. (2026) define J-space through sparse non-negative combinations of token-labelled lens vectors in the *source activation space*. Those vectors live in the residual stream for a residual lens, and in the recurrent-state coordinates for its state analogue. The shared output projection studied here is therefore distinct from an intersection of the two source-space concept dictionaries. Its usefulness for semantic readout must be tested by decoding the projected representations.

Constructing the shared projection. We use the existing conforming lenses, with J_{full} primary and J_{read} as a sensitivity analysis. Separately at every source block of each model, we obtain each lens's leading 64 left singular vectors from its output Gram $JJ^\top$. Principal-angle analysis pairs directions a_i, b_i from these two subspaces, ordered by decreasing non-negative alignment $c_i = a_i^\top b_i$. Their normalised midpoints define an orthogonal projector:

$$q_i = \frac{a_i + b_i}{\sqrt{2 + 2c_i}}, \qquad P_{\ell,k} = \sum_{i=1}^{k} q_i q_i^\top. \tag{13}$$

We predeclare $k = 16$ as primary and $k = 64$ as sensitivity, retain projectors at all 63/63/23 source blocks, and use no evaluation labels to select them. Thus the primary construction takes the 16 most aligned pairs within the two leading 64-dimensional subspaces. Controls use each lens's own leading k output directions and ten seeded Haar-random rank-k projectors. For any lens or local map J, retained energy is

$$E(P, J) = \frac{\|PJ\|_F^2}{\|J\|_F^2}. \tag{14}$$

This measures concentration of derivative sensitivity; it does not measure the number of concepts encoded or recovered.

Shared sensitivity across contexts and delays. Table 29 shows strong alignment of the selected directions and substantial energy retention in both fitted lenses. The selected-pair alignment is higher than the direct overlap between the two own-top-16 subspaces, because it searches within their leading 64 directions.

Table 29: Output geometry of the fitted residual and J_{full} state lenses, averaged over the protocol band. Alignment is the mean squared cosine of the 16 closest principal pairs between the leading 64-dimensional subspaces. Own overlap is the direct top-16 subspace overlap, normalised by 16. Energy columns use the shared rank-16 projector; the random column gives residual-lens energy under rank-16 random projectors.

Model	Alignment	Own overlap	Residual energy	State energy	Random
Mamba-1	0.887	0.500	0.381	0.752	0.0062
Mamba-2	0.882	0.504	0.554	0.659	0.0063
Mamba-3 SISO	0.894	0.593	0.400	0.619	0.0079

On 200 held-out WikiText prompts, we also compare local residual and full-path state derivatives from the same source position to the same final-block target position. We probe four approximately matched depths, at blocks 8/24/40/56 for Mamba-1/2 and 3/9/15/21 for Mamba-3, with lags 0/4/16/64 tokens. Target character boundaries are matched across tokenisers; valid-prompt counts are 200/187/175/97 for Mamba-1/2 and 192/185/170/98 for Mamba-3. Shared-16 captures 12–73% of local-map energy across the measured cells, compared with 0.6–0.8% for random projections. A common component coexists with context dependence: same-context residual/state top-16 overlap at lag zero is 0.34–0.80, while cross-context overlap is 0.10–0.31. For example, Mamba-1 block 8 gives 0.802 within context and 0.119 across contexts. These quantities do not establish a context-invariant semantic format.

Concept recovery within and outside the shared projection. We apply the fixed projectors to clean transported readouts and evaluate all six official task families. Table 30 uses the post-update, normalised joint readout: each component is projected, normalised to unit norm, and then summed before final normalisation and unembedding. The complement replaces $P_{\ell,16}$ with $I - P_{\ell,16}$. These are derived readouts of the existing lenses, with unchanged model activations.

The shared projection has a median relative AUC loss of 66% across the 18 model–family cells (range 29–95%). Nevertheless, its difference from random is positive on 13 cells under item-level paired 95% bootstrap intervals (2000 resamples, without multiplicity correction). Poetry is numerically below random on all three checkpoints. Against own-top-16 joint readouts, the shared projection is lower on five cells, higher on one, and has intervals overlapping zero on twelve. This comparison is specific to the joint readout: for residual-only recovery, shared-16 exceeds residual own-top-16 on fourteen cells under the same interval convention.

Table 30: Normalised joint intermediate-recovery AUC over all source blocks, using post-update residual and J_{full} state readouts. Shared and own retain 16 directions; own projects each component onto its own lens's top-16 subspace. Random averages ten rank-16 projectors. Complement removes the shared directions and retains $d_{\text{model}} - 16$ dimensions; its comparison with full tests removal, rather than equal-rank compression.

Model	Task	Full	Shared	Complement	Own	Random
Mamba-1	Multihop	0.554	0.204	0.534	0.211	0.019
	Multilingual	0.710	0.106	0.666	0.100	0.042
	Order of ops.	0.801	0.389	0.807	0.722	0.106
	Poetry	0.115	0.039	0.113	0.043	0.058
	Association	0.337	0.116	0.274	0.073	0.038
	Typo	0.712	0.150	0.853	0.079	0.066
Mamba-2	Multihop	0.679	0.233	0.643	0.374	0.045
	Multilingual	0.647	0.166	0.640	0.271	0.070
	Order of ops.	0.807	0.428	0.829	0.690	0.265
	Poetry	0.114	0.040	0.106	0.059	0.082
	Association	0.231	0.137	0.195	0.136	0.049
	Typo	0.714	0.152	0.782	0.132	0.088
Mamba-3 SISO	Multihop	0.484	0.094	0.514	0.120	0.025
	Multilingual	0.393	0.156	0.366	0.155	0.028
	Order of ops.	0.518	0.367	0.484	0.432	0.119
	Poetry	0.050	0.003	0.046	0.003	0.029
	Association	0.141	0.037	0.115	0.024	0.025
	Typo	0.560	0.058	0.606	0.051	0.040

Removing the shared directions changes joint AUC by -0.063 to $+0.141$ relative to the full readout. In particular, Mamba-2 multihop changes from 0.679 to 0.643 after removal, whereas retaining only the shared directions gives 0.233. At rank 64, the shared projection retains more recovery, with a median loss of approximately one third; it has no positive difference from own-top-64 joint readouts whose interval excludes zero. Pre-update readouts and removal of the largest pooled sensitivity mode retain the broad geometry–recovery separation. The chosen common directions carry useful signal but provide an incomplete semantic readout. The complement may itself contain shared information, so this result does not identify interface-private concepts or exclude other shared semantic constructions.

Interventions through the projected lenses. We also construct edit directions from $P_{\ell,16}J_\ell$, using the paired two-hop and flexible-generalisation trials, G, X_{unit} and X_{raw} operators, and per-site norm-matched random controls of Section 5. Both state-lens variants and $R/S/RS$ sites are retained; Mamba-1 uses the same sampled suite coverage as that comparison. Projected directions usually give success rates close to random-projector directions. For Mamba-1 flexible generalisation, J_{full} $X_{\text{unit}} : RS$ falls from 30/192 with the original directions to 7/192 with shared-16. A positive Mamba-3 J_{full} $G : RS$ result, 14/192 against 2.7 successes averaged over random-projector families, is matched by own-top-16 at 14/192. These outcomes constrain the tested edit constructions; they do not establish that the shared subspace is unused. Simultaneous RS interventions apply both edits at full strength and do not match the total budget of a single-interface edit.

Numerical qualifications. The local-map pilot's rule that any failed check blocks collection was not applied to Mamba-2's sequential-scan and finite-difference tolerance misses or Mamba-3's finite-difference misses, diagnosed as backend rounding. Mamba-2 state-Gram eigenvalues agree with independently computed earlier dumps to approximately 8×10^{-4}. Mamba-3 local maps differentiate an exact fp32 recurrence whose forward differs from the bf16 kernel by approximately 2×10^{-3}; the kernel backward had discrepancies up to 30% at long lags. The missing BOS token in the initial Mamba-1/2 pilot position plan was corrected before full collection. Under the dimension-scaled fp32 rank tolerance, Mamba-2's fitted J_{full} lens has rank below 16 at 28/63 blocks. Held-out energy retention supports the usefulness of the chosen directions, but the fitted-lens noise floor was not estimated.

Before projection, the residual, J_{full} and joint readouts reproduce all saved protocol ranks under the original decoder. Projection scoring uses the equivalent positive-rescaling-invariant RMSNorm ranking rule; fp32 near-ties change 16/7/6 of 223250/223250/82530 checked rank-capped cells for Mamba-1/2/3. All projected and full baselines in Table 30 use that same scoring path. The results describe these three checkpoints and the specified projectors. They neither demonstrate a natural state-to-residual transfer nor explain Mamba-2's earlier clamp response.

I Implementation and numerical checks

I.1 Constructing and accumulating the state lenses

Derivative paths. Fix a prompt, a source block and an output coordinate j, and suppress the block index. Treat state vectors as columns and derivatives of the scalar $Z_j = \sum_{u \in V_x} z_{u,j}$ as rows. Write

$$r_t = \frac{\partial Z_j}{\partial y_t} C_t, \qquad g_t = \frac{\partial Z_j}{\partial s_t}, \qquad A_{t+1} = \frac{\partial s_{t+1}}{\partial s_t}. \tag{15}$$

Here A_{t+1} is the discrete state transition at the captured inputs, including any coordinate rotation. The scan's input-dependent parameters are fixed when differentiating with respect to its state: the source-block inputs are upstream of that state. The input write and skip term therefore contribute no additional state derivative. All paths leave s_t either through its current output y_t or through s_{t+1}, giving the reverse recurrence

$$g_{T_x-1} = r_{T_x-1}, \qquad g_t = r_t + g_{t+1} A_{t+1} \quad (t = T_x - 2, \dots, 0). \tag{16}$$

The jth rows of J_{read} and J_{full} are respectively the fitting averages of r_t and g_t in Equation 5. The downstream derivative $\partial Z_j / \partial y_t$ is common to both; the control removes only the second term of the recurrence.

Computing the rows. For each prompt, we record the final-block residual and each probed block's scan output before gating and output projection. Backpropagating a unit cotangent in coordinate j at every valid target position simultaneously computes $\partial Z_j / \partial y_t$. Multiplication by C_t gives r_t, and Equation 16 gives g_t. We sum each over valid source positions, divide by $|V_x|$, and accumulate the resulting rows in fp32 on the CPU. After all prompts, division by N_{p} yields the lenses; shards are merged with weights equal to their prompt counts. Output coordinates are processed in batches of size b using replicated prompts, requiring one forward pass and $\lceil d_{\text{model}}/b \rceil$ backward passes per prompt. Residual-lens rows are captured in the same passes as a consistency check. This procedure is implemented in `src/fit_state_lens.py`, with family-specific captures in `src/ssm_states.py` and `src/mamba2_states.py`.

Accumulation without per-token state Jacobians. The implementation evaluates the same source-position sum using forward weights. For an elementwise decay a_t, set $w_{-1} = 0$ and compute

$$w_t = a_t \odot w_{t-1} + \mathbf{1}[t \in V_x], \qquad \sum_{t \in V_x} g_t = \sum_{t=0}^{T_x - 1} w_t \odot r_t. \tag{17}$$

Each weight collects the decay products from all valid earlier source positions. Mamba-1 uses channel- and state-coordinate-specific decays; Mamba-2 uses a scalar decay per head, broadcast over its state coordinates. The weighted readout contractions give the required sum without storing a Jacobian at every token. The readout-only sum instead uses the indicator $\mathbf{1}[t \in V_x]$ directly.

Mamba-3 frames and MIMO readout. Let $\widetilde{s}_t = Q_t s_t^{\text{abs}}$ denote the state in the current token's rotary frame, where Q_t is the corresponding orthogonal rotation. State derivatives transform as $g_t^{\text{cur}} = g_t^{\text{abs}} Q_t^\top$ before averaging. In this frame the transition combines decay with the rotation between successive token frames; the reverse derivative applies its adjoint. The implementation contracts readout vectors with accumulated decay and rotary-phase weights, giving the same sum as the reverse recurrence. It also stores an absolute-frame full-path lens as a separate comparison. For MIMO, y_t includes the readout-rank axis and r_t sums the contractions over that axis; gating and rank reduction remain in the downstream derivative. These operations retain the same state-to-final-residual map dimensions.

I.2 NUMERICAL CHECKS AND VERIFICATION COVERAGE

Exact sequential recurrences validate state capture: Mamba-1's CUDA scan agrees with the torch loop to approximately 1.6×10^{-6}; Mamba-3's kernel agrees with the fp32 reference to 2–4×10^{-3} with bf16 inputs/outputs and fp32 state accumulation. Reverse-scan derivatives were checked against autograd and captured residual rows against the residual-lens estimator. At the 23 blocks shared by Mamba-2's sparse and 63-block runs, per-block recovery AUC agrees to four decimal places.

The per-position release records post-update residual, state and joint ranks at every probed block and prompt position: 541 scored prompts each for Mamba-1/2 and 545 for Mamba-3 SISO, including 93 multihop prompts each. Association token coverage accounts for the difference (92/92/96 of 102). The selected temporal maps remain illustrative; population span-persistence statistics are not reported here.

Verification coverage. Table 31 summarises automated consistency checks on the archived results, covering summary values, result records, rank-array checksums and arithmetic. These checks validate the covered numerical artefacts; they do not independently replicate the model runs. Separately, all 72 recovery AUC values for the conforming Pythia baseline and the Mamba-2 row-zeroed lens were reconstructed from per-item ranks: six families, three lenses, two layer restrictions and two runs.

Paired verbal-report counts at top-1 and top-5 were reconstructed from the per-trial records, including checks of trial pairing and set unions.

Table 31: Numerical consistency checks on the archived results. All listed checks passed. Counts refer to assertions, not independent experiments.

Analysis checked	Assertions
State and joint readouts	974
Held-out residual Jacobian geometry	752
Held-out state Jacobian geometry	4077
Readout protocols and intervention controls	2214
Direction and fixed-budget controls	508
Coordinate exchange and numerical cross-checks	1234
Paired coordinate exchange and sign-guarded steering	3030
Shared output projections and concept recovery	2746
Paired verbal reports: Mamba checkpoints	782
Paired verbal reports: Nemotron and Pythia	57
Workspace task suite	4347
Pursuit interventions	1138

Three audits of saved model outputs reproduce their archived summaries exactly. The complete two-hop audit checks 450 trial records and their evaluation metadata; separate two-hop and flexible-generalisation audits verify edited-concept outputs. Provenance records link tables and figures to source artefacts, preserving SHA-256 hashes, denominators, scoring definitions and the Mamba-3 replay qualifications. Verification is limited to the available archived artefacts and does not establish reproducibility of every historical experiment.