LOOPED TRANSFORMERS UNDER THE JACOBIAN LENS: DOES THE GLOBAL WORKSPACE SURVIVE RECURRENCE?

Anonymous authors
Paper under double-blind review

ABSTRACT

A Jacobian lens identifies verbalisable, causally editable representations in standard transformers, but it is unclear whether the same workspace functions arise when depth reuses weights. We extend the lens to recurrent architectures by virtual unrolling and fit distance-controlled lens families for Ouro-2.6B and Huginn-0125, with Qwen3.6-27B as an untied baseline. Readouts, sparse concept inventories and eleven behavioural experiment families reveal two maintenance regimes. Ouro reconstructs content across supervised loop checkpoints: sustained writes outperform single-loop writes, while recovery from ablation depends on its width. Huginn's content remains readable across sixteen recurrences, but input-path interventions show that this persistence depends on continued re-injection; reads and effective edits follow a roughly two-recurrence horizon. Reporting existing and newly injected content also dissociates: Huginn excels at the former yet fails at the latter, which only Ouro achieves among the tested checkpoints. The results support functional workspace organisation under recurrence while distinguishing model behaviour from limitations of the readout instrument.

1 INTRODUCTION

Global workspace theory links flexible behaviour to representations that are widely available, reportable and open to directed modulation (Baars, 1988; Dehaene & Changeux, 2011; Mashour et al., 2020). Gurnee et al. (2026) operationalise these properties in language models using the *Jacobian lens*: linear maps derived from the model's own mean input–output Jacobians. In a standard transformer, they identify a mid-depth band whose content can be verbalised, causally edited and selectively engaged by task demands. Their account leaves an architectural question open. Biological workspace theories emphasise recurrent processing, whereas an untied transformer visits each layer once. Does the same functional organisation arise when depth is implemented by re-applying the same weights?

We study two released weight-tied models with different iteration interfaces: Ouro-2.6B, which loops its full stack and supervises every loop end, and Huginn-0125, which recurs a small core without explicit intermediate supervision. We reproduce the untied Qwen3.6-27B baseline under the same harness. A direct application of the final-target lens would give a misleading answer: poor transport across iterations can conceal content that is present locally. We therefore fit and validate a family of lenses targeting different iterations before comparing readouts and interventions.

Both recurrent models exhibit verbalisable, causally editable and selectively engaged representations, but their access and maintenance differ. Ouro reconstructs content across its supervised checkpoints; Huginn's content remains readable across recurrences while depending on repeated input injection. Neither observation alone establishes an autonomous memory buffer. Our contributions are:

- **Validated lenses for recurrent depth.** Virtual unrolling exposes each firing of a tied block to the original fitting procedure. Distance-controlled lens families separate failed transport from absent content (§§3–4).

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.

- **Two maintenance regimes.** Readout, sparse concept inventories and input-path interventions distinguish repeated reconstruction on Ouro from input-dependent persistence on Huginn (§5).
- **Causal interfaces.** Writes and ablations reveal different intervention horizons, with a crucial dependence on ablation width. Steering existing content also dissociates from reporting newly injected content (§6).

The conclusions concern these models' measured functions. We make no claim about consciousness or an isolated causal effect of training architecture.

2 Background and related work

A *looped transformer* re-applies a weight-tied block within one forward pass, following the Universal Transformer and parameter-sharing models (Dehghani et al., 2019; Lan et al., 2020). Ouro loops its entire stack; Huginn recurs only a four-layer core between an untied prelude and coda (Table 1). Ouro explicitly trains the language-model head at every loop end. Huginn trains the post-coda output with sampled recurrence counts and truncated backpropagation, re-injecting the prelude encoding at each recurrence (Zhu et al., 2025; Geiping et al., 2025). We use 16 recurrences. Its virtual depth is close to Qwen's, but parameter count and computation per layer differ; this is not a compute-matched comparison.

Table 1: Models and iteration interfaces. Virtual depth counts block applications per forward pass. Training recipe and scale are not controlled.

Model	Parameters	Tied computation	Depth	Intermediate supervision
Qwen3.6-27B	27B	64 untied layers	64	None
Ouro-2.6B	2.6B	48 layers $\times 4$	192	Every loop end
Huginn-0125	3.5B	$2 + (4 \times 16) + 2$	68	No explicit per-step loss

Related work. Workspace theory motivates functional tests of information availability; its use in AI does not by itself establish consciousness (Butlin et al., 2023; Gurnee et al., 2026). Readout methods differ in their maps: the logit lens directly unembeds intermediate states (nostalgebraist, 2020), the tuned lens learns an affine decoder (Belrose et al., 2023), and the Jacobian lens estimates local transport from the model. Its failure can therefore reveal limits of the measurement as well as properties of the computation. On Huginn, Lu et al. (2025) find that logit and coda lenses disagree systematically across core blocks, with little evidence of an interpretable latent chain of thought. Blayney et al. (2026) study iteration dynamics through hidden-state geometry, attention patterns and linear probes. These approaches characterise recurrent states but do not supply the between-layer transport maps used by the workspace suite. We combine those maps with concept-specific interventions to ask whether readable content also affects behaviour. Extended positioning, including work on iterated algorithms (Giannou et al., 2023; Yang et al., 2024), is in Appendix N.

3 Method: a lens family on the unrolled graph

For source state h_v and downstream target h_t, written as column vectors, the Jacobian lens is the corpus-mean Jacobian of the intervening computation (Gurnee et al., 2026):

$$J_{v \to t} = \mathbb{E}\left[\frac{\partial h_t}{\partial h_v}\right]. \tag{1}$$

Readout unembeds $J_{v \to t} h_v$. A token with unembedding vector u supplies a unit intervention direction $d = J_{v \to t}^{\top} u / \|J_{v \to t}^{\top} u\|$. We add that direction, remove its projection, or clamp-transfer projections between token directions. Readability is assessed by probe-token ranks and concept inventories; causal edits test whether the inferred directions affect behaviour. We abbreviate $J_{v \to t}$ to J_v when the target is understood.

Virtual unrolling. Each firing of a physical block becomes a distinct *virtual layer*. With zero-based indices, Ouro uses $v = 48s + i$ for loop s and layer i; Huginn's core uses $v = 2 + 4s + i$, with prelude and coda blocks firing once. Virtual-block proxies dispatch recording hooks only on their assigned firing, preventing a standard recorder from silently retaining only the last iteration. Automatic differentiation follows the actual unrolled graph, including its normalisation and tied weights. Per-pass derivatives are exact for that graph; their corpus mean remains an estimated linear summary. Validation checks distinct per-firing activations, agreement with the model's own final logits, non-zero gradients and deterministic forwards. It also catches no-grad recurrence phases, random initialisation and checkpointing that re-fires hooks (Appendix G).

Fitting and target selection. All models use the same 1,000 WikiText-103 prompts of 128 tokens, excluding the first 16 positions, with frozen parameters and the original mean-Jacobian estimator. Fits and experiments use bf16 on one to three NVIDIA B200 GPUs; source/target coverage, convergence and timings are in Appendix I. A final-target lens alone is insufficient: it reads punctuation through Ouro's first three loops, while Huginn's useful transport extends only about two recurrences before its target. These failures do not establish that the intervening states lack content.

We therefore target each Ouro loop end (v47/95/143/191). For Huginn, the original four lenses target recurrence ends at $r = 4/8/12$ and the final coda at $r = 16$; these support the causal suite. Additional fits complete the *readout* family at every remaining recurrence end. We compare sources at matched offsets from their own targets and use only locally validated readouts. Completing the tiling prevents unequal lens distances from creating spurious depth trends. Probe-token recovery and concept-specific interventions validate useful semantic directions; whole-state norm or output agreement alone cannot validate a lens at range, as the next section shows.

The target choice also determines the interpretation of a readout. Ouro's loop ends are precisely the states its intermediate losses train the head to read, so local transport terminates at a trained decoding interface. Huginn has no equivalent intermediate loss: we check its local maps empirically, rather than assuming that every core state is a calibrated decoder. The readout and intervention operators share the same pulled-back token directions. A direction that both recovers a prompt-absent concept and changes a concept-dependent answer supplies stronger evidence than a plausible token list alone.

4 Transport and readable geometry

4.1 Semantic transport differs from bulk transport

We distinguish transport of the residual state as a whole from recovery of a particular concept. Norm, state cosine and output-distribution agreement summarise aggregate behaviour; a prompt-absent probe token tests whether a specific semantic direction survives. The two need not agree, because useful concept directions may occupy only a small part of the state.

A mean Jacobian retains directions shared across prompts; prompt-dependent responses can cancel in the average. Whole-state transport consequently weakens with distance in all three models. At a span of 39 layers, Qwen's transport has only 2% top-1 agreement with the final output and normalised norm 0.46, within Ouro's 0.14–0.53 range at matched distance. Nevertheless, Qwen's same map still carries a specific latent concept. Poor aggregate agreement therefore does not imply that every semantic direction is lost.

We separate these effects with the prompt "Fact: The currency used in the country shaped like a boot is…". The intermediate concept *Italy* is absent from the prompt but independently readable at both endpoints of each transport in Table 2. It survives Qwen's within-stack map and Ouro's within-loop maps. Across an Ouro loop boundary, however, its projection falls to approximately 0.1 of the target-state value, without concept-versus-control discrimination, while bulk magnitude retains 0.27–0.46. Thus, the cross-boundary failure concerns the tested semantic component even when whole-state alignment is better than in Qwen's working case. It does not establish that all information is destroyed in the model.

Ouro's loop ends are supervised *checkpoints*: states trained to be readable by the language-model head, rather than saved parameter files. Direct unembedding at these checkpoints agrees with the final output on 60–92% of tokens. Together with the local readouts, the boundary result supports

Table 2: Transport of the prompt-absent concept *Italy*. Both endpoints contain the concept; the last column tests recovery after linear transport. Better state cosine does not guarantee semantic recovery.

Transport	Distance	State cosine	Italy recovered?
Qwen: band to target	39	0.19	Yes
Ouro: band to own loop end	11–23	0.3–0.5	Yes
Ouro: loop-2 band to loop-3 band	32–60	0.31–0.45	No

a decode/re-encode account: successive loops exchange a token-aligned state and reconstruct their own workspace representations. This interpretation is tested behaviourally in §6.1; attributing it specifically to deep supervision would require controlled training comparisons.

4.2 Huginn's readable window follows the target

Huginn has no comparable fixed workspace band. On "Fact: The capital city of the country famous for sushi is", *Tokyo* reaches rank 1 at recurrence steps 2–3 through the $r = 4$ lens, step 7 through $r = 8$, steps 10–11 through $r = 12$, and steps 14–15 through the final-target lens (Figure 1). Every lens reads the same 16-recurrence forward pass, but each resolves content only in the one-to-two recurrences before its target. Failure of an early source to reach a distant target is therefore an instrument limit, not evidence that early states are empty.

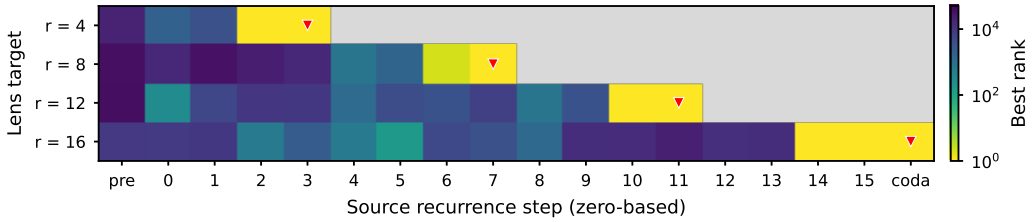

Figure 1: Huginn on the sushi prompt: best candidate rank, bucketed by source recurrence, for four lenses reading the same forward pass. Red triangles mark targets; grey cells are outside fitted source ranges. Low ranks occur near each target. Full three-prompt figure: Appendix K.

The fitted maps also repeat at matched offsets: independent Huginn lenses with $r = 8$ and $r = 12$ targets have matrix cosine 0.9991, versus 0.43 when misaligned by one recurrence. This near-invariance supports comparisons in iteration-relative coordinates. Symmetry and mean-transport factorisation analyses are reported in Appendix J; they are empirical similarities, not guarantees of exact lens transfer.

Mean transport also differs in how it composes through a cut. The cosine between a directly fitted map and the product of two fitted maps is 0.94–1.00 at Ouro's supervised loop ends, 0.69–0.95 at Qwen's band entry, and 0.25–0.84 at Huginn's recurrence ends. Averaging breaks the per-prompt chain-rule identity when responses on either side of the cut are correlated. The observed ordering therefore adds a structural comparison, but does not by itself prove a Markov property or information loss at a checkpoint. We next examine the evolving content and test its maintenance directly.

5 Maintenance and occupancy across iterations

The validated lens family lets us distinguish when a concept becomes readable from how later computation maintains it. Readouts are compared at matched source-to-target offsets throughout.

5.1 Ouro reconstructs content across loops

On the boot riddle, Ouro's first-loop band contains literal associations (*boots, footwear, shoe*); the second-loop band already reads *Italy*, which remains readable in later loops. Combined with the failed

cross-boundary transport, this supports reconstruction in each loop's coordinates. Other probes settle at different depths: an easy capital-city answer is top-1 at the first checkpoint, while an arithmetic answer settles in the band only in loop 4. Over WikiText, loop 3 already agrees with the final output on 92% of tokens. Checkpoint readouts and the model's trained early-exit gate provide checks independent of fitted transport (Appendix K). The causal tests below distinguish recovery after an early edit from resistance to a late one.

The three timing measurements have different dependencies. Per-loop ranks require fitted maps; direct checkpoint unembedding uses only the trained language-model head; and the early-exit gate is a component of Ouro itself. Its mean exit probabilities across loops are $(0.03, 0.14, 0.34, 0.49)$. The gate often favours later computation even when the output token is already stable, so token agreement should not be read as proof that the final loop performs no useful computation. Together these measurements distinguish answer stability from the internal work performed after it is first found.

5.2 Huginn's persistence depends on re-injected input

On Huginn, the same latent answer is readable through successive local windows from recurrence 2–3 to the end of the pass. This apparent persistence could reflect a stored state or continued reconstruction from the prelude encoding, which is re-injected at every recurrence. Two interventions on that injection path leave the incoming recurrent state untouched.

First, scaling the injected encoding to at most one quarter of its magnitude prevents recovery of the baseline next token on all eight prompts, even when only the last four recurrences are affected. Removing it from recurrence 8 raises the latent answer's median rank from 11 before the cut to 2,749 after one recurrence, then roughly 40,000. Second, swapping in the encoding of a matched-length prompt changes the final output to the new prompt's answer in all 30 tested configurations, including late swaps. The old answer fades within roughly two recurrences (Appendix L). These results support continuous, input-dependent recomputation under this protocol rather than autonomous retention of a completed answer. They do not isolate the benefit of input injection as a training design choice.

5.3 A concept inventory recovers the depth profiles

We cross-check targeted ranks using the original study's demonstration protocol: eight chat prompts, 128-token greedy completions and recorded (layer, position) states. Our implementation of its concept inventory uses gradient pursuit (Blumensath & Davies, 2008; Nanda et al., 2024) to reconstruct each state from unit-normalised atoms $a_\tau = J_{v\to t}^\top u_\tau / \|J_{v\to t}^\top u_\tau\|$. A token occupies a cell when it is among the first 25 atoms and its fitted coefficient exceeds $0.05\|h\|$. Counts require no target answer; Appendix E specifies the algorithm and the full parameter grid.

Figure 2 uses the animal prompt ("I'm thinking of the largest land animal. What do they eat? Answer briefly."), the only demonstration prompt all three models answer directly and correctly. Tokens are classified as *latent* when absent from the transcript and *spoken* when present, merging singular/plural variants. Qwen's latent counts enter a mid-depth band; Ouro shows a band in every loop. On Ouro, in-band occupancy exceeds the out-of-band *per-layer rate* in all four loops at the primary setting (loop 4: 22.3 versus 16.6). Raw totals obscure this difference because the out-of-band region contains more layers.

Completing Huginn's readout tiling holds target distance constant across recurrences. Window totals then plateau at 31–44 slots over v17–v60, before jumping to 92 in the coda window. The earlier partial tiling instead produced an apparent increase with depth. That increase reflects lens distance, not accumulating content (Appendix D). Qualitative depth structure also appears under top-8 readout, but token selections and some conclusions differ. In particular, Qwen's apparent zeros depend on coefficient thresholds, and Ouro's band contrast weakens at 50 atoms. Counts describe this instrument's recovery, not an exhaustive inventory of the state.

This qualification is consequential. The pursuit support and top-8 tokens have mean cell-level Jaccard overlap only 0.07–0.24, yet recover the same broad depth profiles. Two finer top-8 conclusions do not survive pursuit: a late thinning of latent relative to spoken tokens on the animal prompt, and an Ouro band advantage based on unnormalised loop totals. At 50 atoms, pursuit explains only 4.3% of Qwen's state variance, 10.2% of Huginn's and 21.5% of Ouro's. Agreement between instruments

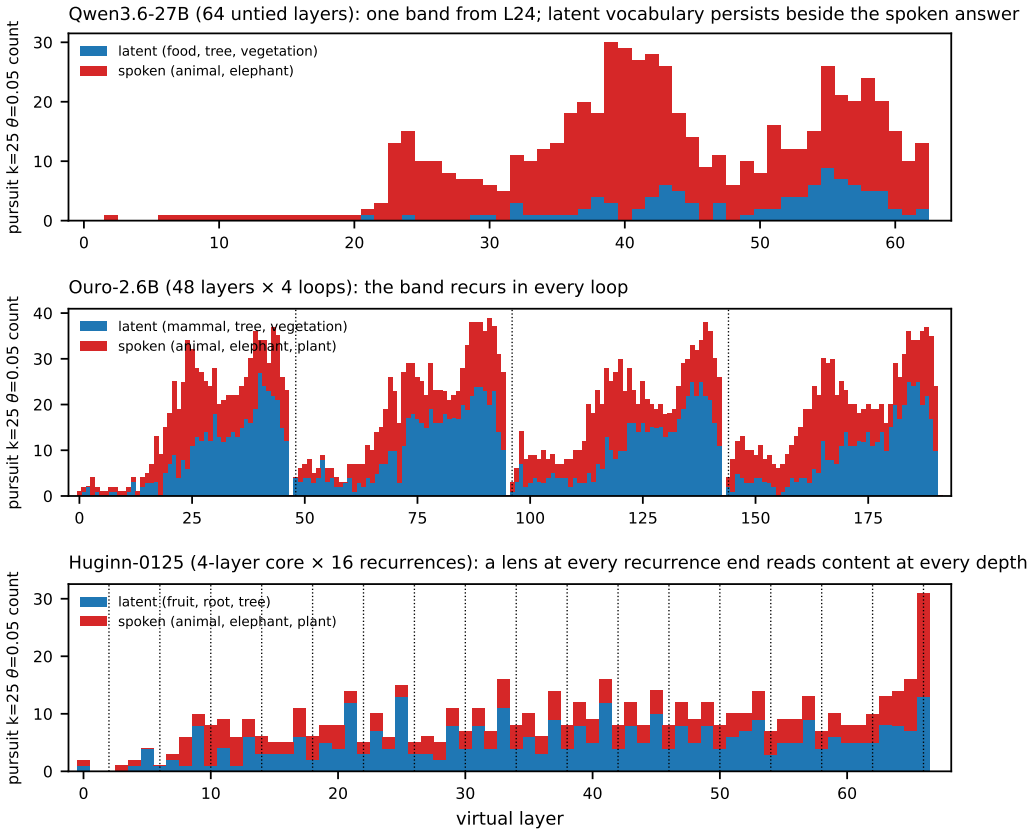

Figure 2: Pursuit occupancy on the animal prompt, with 25 atoms and coefficient threshold $0.05\|h\|$. Blue latent and red spoken counts are stacked. Dotted lines mark iteration boundaries. Ouro's bands repeat within loops; Huginn's local sawtooth follows distance to each lens target. Its window totals plateau after early recurrences. Transcripts and estimator controls: Appendices D–E.

strengthens the geometric comparison; their disagreements bound what can be claimed about the full representation.

The inventory also exposes changes within the repeated bands. On the riddle asking which currency the boot-shaped country used before 2002, Ouro answers "the Italian lira". The unspoken alternative *euro* peaks in loop 2 and then declines from 259 to 188 occupied slots, while *Italy*, the intermediate concept needed for the correct answer, rises from 177 to 390. This shift survives all twelve estimator settings. It shows how similar band-level occupancy can coexist with a changing balance between candidate concepts. Token classes nevertheless remain incomplete: Qwen's inventory also includes a Chinese token for Italy (102 slots, alongside 344 for English *Italy*), so English-only counts miss part of the recovered concept.

Direct decodability. The logit lens agrees with Huginn's emitted next token at 15–25% of positions across the recurrent stack, compared with 1–7% in Ouro's mid-loop states and at most 9% in Qwen's L8–40. Ouro instead has sharp checkpoint peaks. Huginn's sampled training depths may exert indirect pressure towards decodability, but weak direct readability does not imply explicit deep supervision.

The probability field gives the same ordering with a different score. The mean probability assigned to the actually emitted next token is 0.61–0.65 at Ouro checkpoints, 0.02–0.045 through Huginn's recurrent core, and 0.47 at Huginn's coda. These are lower bounds from top-8-truncated dumps. On Qwen at L36, direct unembedding gives 0.003, versus 0.019 after Jacobian transport. Thus, a concept can be useful to a model without its intermediate state being a high-confidence next-token

prediction. This matters when interpreting the different report tasks below. Full probability profiles and secondary prompts are in Appendix D. The next section tests whether readable content is also accessible to causal edits.

6 CAUSAL ACCESS TO WORKSPACE CONTENT

We run the eleven behavioural experiment families of Gurnee et al. (2026) in one shared harness. They include edits along lens directions and readout probes of task selectivity and capacity. Qwen's intervention interface is L24–48; Ouro uses the bands of loops 2–4; Huginn uses its last three validated lens windows. Unless specified otherwise, edits act at every token position and every listed intervention layer. Introspection instead injects over the user-instruction span. Table 3 isolates the conditions that test the interfaces; Appendix B reports the full suite, token-position conventions and exact counts.

Table 3: Causal interfaces. Write cells count successful flips out of 192; ablation cells count previously correct answers retained (Qwen $n = 36$, Ouro $n = 42$, Huginn $n = 28$). "Early" means Ouro's loop-2 band or Huginn's second-to-last window; "last" means the final interface. Qwen has one interface. Verbal-report entries are top-5 rates; its Qwen result is a stored aggregate. Introspection entries count top-1 reports at 8σ.

Outcome / condition	Qwen	Ouro	Huginn
Verbal report: existing content	69%	106/133 (80%)	124/126 (98%)
Introspection: new content	0/101	31/101 (31%)	0/97
Write: all interfaces	80/192	73/192	25/192
Write: early only	—	40/192	5/192
Write: last only	—	48/192	25/192
Argument ablation: all	26/36	30/42	17/28
Argument ablation: early only	—	42/42	28/28
Argument ablation: last only	—	32/42	17/28
Unrelated-direction control: all	36/36	42/42	28/28
Ten-direction ablation: all	23/36	0/42	1/28
Ten-direction ablation: early only	—	6/42	27/28
Ten-direction ablation: last only	—	0/42	0/28
Random-atom subspace control: all	35/36	37/42	26/28

6.1 WRITES DEPEND ON WHEN THE INTERFACE IS EDITED

The flexible-generalisation task swaps a concept and scores whether the top-1 next token changes to the answer implied by its replacement. For example, on "The capital of France is the city of", a France-to-Canada intervention succeeds when the model predicts *Ottawa*. At each selected state we apply clamp-transfer:

$$h \leftarrow h - (h \cdot d_A)d_A + s(h \cdot d_A)d_B, \tag{2}$$

where d_A and d_B are the original and replacement concept's unit lens directions, and $s = 1$ unless specified. This tests a consequence of the substituted concept; verbal report instead scores the replacement token itself.

Qwen flips 42% of 192 trials, close to the original study's 40%. On Ouro, clamping all remaining loop bands reaches 38%; clamping loop 2, 3 or 4 alone reaches 21%, 29% or 25%, respectively. The early and late failures differ. Under loop-2-only clamping, failed trials show the replacement answer reaching median rank 7 at that checkpoint but falling to 44 by loop 4 while the original answer recovers to rank 1. Under loop-4-only clamping, the original answer is already rank 1 at entry and the replacement rises only to median rank 11. The first pattern indicates recovery after an effective edit; the second indicates resistance to displacement, with no later loop available to undo it. Sustained writes counter both effects. Full trajectories and category splits are in Appendix M.

On Huginn, the last window alone produces the same 25/192 flips as all three windows, including identical per-category counts; the second-to-last window alone produces 5/192. Its short readout

horizon therefore also predicts the useful range of these writes. Comparing success rates across models requires care: no-intervention task accuracy is only 28/64 on Huginn, versus 42/64 on Ouro. The stronger within-model evidence is the change in effect when the same intervention moves between interfaces.

Intervention magnitude does not substitute for placement. Doubling the clamp strength reduces all-interface flips to 14/192 on Ouro and 2/192 on Huginn, from 73/192 and 25/192 at the main setting. A larger perturbation can therefore disrupt the computation rather than enforce the intended concept. The interface comparisons hold the operator and strength fixed; they are not claims that any sufficiently strong edit will follow the same pattern.

6.2 Ablation width qualifies reconstruction

Single-direction ablation removes the projection onto a concept direction, $h \leftarrow h-(h\cdot d)d$. Removing an argument concept impairs all three models, while unrelated-direction controls retain every answer. On Ouro, the early-only argument ablation is a complete null (42/42 retained), consistent with later reconstruction. On Huginn, the last-only and all-window ablations each retain 17/28, whereas the earlier window retains 28/28. These contrasts mirror the write conditions.

A wider intervention changes the Ouro conclusion. We orthonormalise the ten pursuit atoms selected from an uninstrumented forward pass, freeze their span, and remove the state's projection onto it. A matched random-token subspace control retains 87–97% of answers. Ouro's loop-2-only pursuit ablation now retains just 6/42 (14%), against 42/42 for one direction: later loops can recover the narrow deletion but not this broader disruption. Recovery after an early ablation is therefore specific to the width of the removed component. Huginn's last-window and all-window effects remain near-equivalent (0/28 and 1/28 retained), while the earlier window retains 27/28. Qwen retains 23/36 under the wider ablation, compared with 35/36 under its control. These results support an interface constraint on Huginn and a width-dependent recovery process on Ouro, rather than a universal rule about recurrence.

Removing the answer token's direction gives a further contrast: Qwen retains 19/36 correct answers (53%), Ouro 8/42 (19%) and Huginn 0/28. This ordering is consistent with the greater direct decodability of Huginn's recurrent states and Ouro's checkpoints, while Qwen's mid-stack representation is less aligned with its immediate output. The argument-direction and unrelated-word controls are needed to distinguish this output disruption from a concept-specific intervention.

6.3 Existing and newly injected content dissociate

Verbal report reroutes a concept the model has already computed and is about to emit. Huginn reports the replacement in its top-5 on 124/126 trials (98%). Introspection instead injects a concept absent from the current computation and asks the model to name it. No Huginn concept reaches even rank 100 across 97 concepts and the tested strengths, despite its median concept rank improving from 40,666 to 16,258 at 8σ. The perturbation measurably changes the output distribution without making the concept reportable under this task. Its failure therefore concerns successful naming, rather than an absence of any response to the injection. Ouro reports newly injected concepts at top-1 on 31% of trials; Qwen also has a null, with its best injected concept reaching rank 173.

Thus, the two report tasks measure distinct forms of access. In these checkpoints, success on the new-content task coincides with Ouro's explicit per-iteration supervision, whereas Huginn's weak direct decodability coexists with strong steering of existing content. The comparison does not establish that explicit supervision is necessary or sufficient: training, scale and iteration interfaces differ simultaneously.

6.4 Selectivity, nulls and robustness

Selective engagement supplies a further workspace signature. On Huginn, a passage's language is more readable when relevant to the task (hit@5 1.00 versus 0.62), yet the model answers the task's author question correctly on 0/8 prompts. Readable, task-gated content is therefore not sufficient for successful behaviour. This capability floor also qualifies its weaker probe-swap and flexible-generalisation results.

Other failures concern measurement. The capacity task asks the model to hold 20-word lists in mind and probes their contents at comma positions. Replacing rank readout with pursuit at a coefficient threshold of $0.01\|h\|$ raises the measured plateaux from 0.7 to 1.9 words on Qwen, 2.4 to 4.9 on Ouro, and 1.8 to 3.6 on Huginn. At the stricter $0.05\|h\|$ threshold, Qwen falls to 0.1 words (Appendix E). The same states therefore yield substantially different estimates under different recovery criteria. These figures bound what the instruments extract, rather than measuring the models' total memory.

Top-down summoning instead names a category and asks the model to bring an absent concept to mind. All models score near zero, without discriminating summoned concepts from matched foils. This null admits more than one explanation: the requested computation may not occur, or the resulting content may be poorly recovered by the lens. A direction annihilated by mean transport is invisible to that instrument, even if the state contains it. This does not imply that every concept irrelevant to the next token lies in that nullspace (Appendix M).

Robustness to reasoning SFT. We repeat lens fitting, structural analysis and the causal suite on Ouro-2.6B-Thinking. The main interface contrasts recur: sustained writes outperform single-loop writes, newly injected concepts are again top-1 reportable on 31% of trials, and early narrow ablation is nearly harmless while early ten-direction ablation retains only 11% of answers (base: 14%). Other signatures do change. The language-selectivity gap falls from 0.25 to 0.00, and ignition's 10–90% transition width narrows from 0.55 to 0.20 (Appendix C). Thus, this training comparison supports the stability of the intervention interfaces, while showing that selective engagement and transition sharpness can vary within the same architecture.

7 Discussion and conclusion

Weight tying is compatible with the tested workspace functions: both recurrent models contain representations that are verbalisable, selectively engaged and causally editable. Their interfaces differ. Ouro's checkpointed loops reconstruct content after some early edits, although wider subspace deletions can prevent recovery. Huginn's readable content persists across depth only with continued input re-injection, and effective edits remain local to a short lens horizon. Steering computed content also separates from reporting newly injected concepts. These distinctions become visible only after controlling the readout's source-to-target distance.

The scope is limited by both the models and the instrument. One base checkpoint per family confounds architecture with scale and training; an Ouro SFT comparison does not resolve those confounds. Prompt-specific readouts and small behavioural suites do not establish universal workspace organisation. Mean Jacobians capture only shared linear response directions; single-token concept vocabularies, coefficient thresholds and ablation width change what is recovered or disrupted. A failed lens is not evidence of an empty state, and agreement with selected workspace criteria is not a claim about consciousness.

The maintenance results also qualify the idea that later iterations simply store an already completed answer. Ouro can restore content after an early edit, while Huginn remains dependent on refreshed input even late in its pass. Continued computation could support useful self-correction, or it could reflect work that a different iteration interface would avoid. The present experiments do not distinguish these possibilities: recovering an original answer after a deliberately misleading edit is not yet evidence of correcting the model's own reasoning errors. Controlled training runs varying supervision and input injection separately, together with recurrence-count sweeps, could test which boundary designs support persistent, controllable representations. The present results supply measurement controls and causal contrasts for that comparison.

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 A. The authors are responsible for the final text, claims, code and artefacts.

REPRODUCIBILITY STATEMENT

Appendix I specifies the lens-fitting protocol and source/target coverage; Appendices B, D and E describe the behavioural and readout protocols. Adapter validation and compatibility details are in Appendices G–H. Appendix A describes automated checks against the recorded experiment artefacts and identifies families available only as aggregates.

REFERENCES

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

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112v1*, 2023. URL https://arxiv.org/abs/2303.08112v1.

Hugh Blayney, Álvaro Arroyo, Johan Obando-Ceron, Pablo Samuel Castro, Aaron Courville, Michael M. Bronstein, and Xiaowen Dong. A mechanistic analysis of looped reasoning language models. *arXiv preprint arXiv:2604.11791*, 2026.

Thomas Blumensath and Mike E Davies. Gradient pursuits. *IEEE Transactions on Signal Processing*, 56(6):2370–2382, 2008.

Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, et al. Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023.

Stanislas Dehaene and Jean-Pierre Changeux. Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2):200–227, 2011.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1380. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3b01972cf31e6fa0fe29e4b8b5c2a0a1-Abstract-Conference.html.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D. Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. In *International Conference on Machine Learning*, 2023.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *arXiv preprint arXiv:2607.15495*, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

Wenquan Lu, Yuechuan Yang, Kyle Lee, Yanshu Li, and Enqi Liu. Latent chain-of-thought? decoding the depth-recurrent transformer. *arXiv preprint arXiv:2507.02199*, 2025.

George A. Mashour, Pieter Roelfsema, Jean-Pierre Changeux, and Stanislas Dehaene. Conscious processing and the global neuronal workspace hypothesis. *Neuron*, 105(5):776–798, 2020.

Neel Nanda, Arthur Conmy, Lewis Smith, Senthooran Rajamanoharan, Tom Lieberum, János Kramár, and Vikrant Varma. [full post] progress update #1 from the GDM mech interp team. AI Alignment Forum, April 2024. URL https://www.alignmentforum.org/posts/C5KAZQib3bzzpeyrg/progress-update-1-from-the-gdm-mech-interp-team-full-update.

nostalgebraist. Interpreting GPT: the logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Liu Yang, Kangwook Lee, Robert Nowak, and Dimitris Papailiopoulos. Looped transformers are better at learning learning algorithms. In *International Conference on Learning Representations*, 2024.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, et al. Scaling latent reasoning via looped language models. *arXiv preprint arXiv:2510.25741*, 2025.

Table 4: The eleven-family suite, including the original study's reported ranges where available. Metrics differ between families; rates should be interpreted with their task baselines and readout limits.

Family	Study	Qwen3.6-27B 27B	Ouro-2.6B 2.6B	Huginn-0125 3.5B
Verbal report (swap → top-5)	59–88%	69%	80%	98%
Introspection (injected concept)	reportable	null	top-1 31%	null (0/97)
Modulation (focus/ctrl/suppress)	ordered	.38/.07/.07	.77/.39/.26	.28/.15/.19
Top-down summoning	evokes concept	∼0	∼0	∼0
Two-hop probe swap	54–70%	52% (base 63%)	29% (base 48%)	6% (base 28%)
Flexible generalisation (swap)	40%	42%	38%	13% (base 44%)
Selectivity: language	expl. ≫ auto	+1.00	+0.25	+0.38
Selectivity: line count	iff relevant	0.09@1	1.00@1	0.09@1
Ignition (10–90 width)	sharp	0.10	0.55	none in-window
Capacity (words held)	∼25	0.7	2.4	1.8
Dual task (concepts survive)	survive	6/21	17/21	7–8/21

A VERIFICATION OF REPORTED NUMBERS

Reported quantitative results are checked against the recorded raw artefacts — per-trial result records, run logs, and the readout dumps — by an automated verification suite included with the code release (232 checks, plus a companion suite of 132 checks covering the gradient-pursuit artefacts of Appendix E). Each suite runs with one command and checks the corresponding reported statistics. Most values are recomputed from per-trial records, with percentages and 95% Wilson intervals recomputed from raw counts. For four experiment families the artefacts store only condition-level aggregates (directed modulation, dual task, line-count selectivity, and the Qwen3.6-27B verbal report — marked AGG in the tables below); for these the suite verifies the reported value against the stored aggregate but cannot independently re-derive it. Per-trial recording would enable independent re-derivation for those families as well.

B FULL CAUSAL-SUITE TABLES

Tables 5–7 expand the condensed Table 4 to exact counts. Every value is re-derived by the verification suite of Appendix A; rows whose artefacts store only aggregates are marked AGG.

Token-position conventions. The lens itself carries no position index — J_v is a mean Jacobian, marginalised over the fitting corpus's prompts and positions (§3) — so position enters only at readout and intervention time. Our default, matching the original study's stated protocol (Gurnee et al., 2026), is that a direction intervention applies at *every token position* as well as at every listed virtual layer: when §6 says Ouro clamps the bands of loops 2–4, the clamp acts on the whole sequence at every band layer of every remaining loop, and the behavioural outcome is read from the final position's next-token distribution. Verbal-report swaps, flexible-generalisation swaps, two-hop probe swaps, and ablation all follow this default. Two families are span-restricted: the summoning clamp-transfer edits only the stimulus's token span — more conservative than the original all-positions protocol, since the edit reaches the answer position only through attention — and introspection injects over the token span of the user instruction, matching the original's user-turn injection, with the effect read at the final position. Rank-probe readouts (summoning, modulation, selectivity, capacity, dual task) take the minimum lens rank over band layers × a task-defined span: the stimulus or carrier-sentence tokens, every position after the probed passage, or the comma positions of the capacity list.

Two rows deserve notes. Summoning's zero is a *no-discrimination* zero in the strongest sense: summoned concepts score no better than matched foils (difference ≤ 0.01 on all models) and asking versus not asking changes nothing ($|q_1 - q_2| = 0.000$ everywhere). And the line-count selectivity cells read "band readability of the count when directly asked / when never asked": only Ouro shows the gated pattern.

The fourth category, numbers, is omitted from the table for width: it never flips under any condition on either looped model (0/48 throughout; direction cosine 0.64 on Ouro-2.6B). Note the Huginn

Table 5: The full causal suite with exact counts. Columns follow Table 4; "base" is the no-intervention task baseline. Modulation cells are topic hit@5 under focus/control/suppress instructions; capacity cells give the last-quartile plateau with the peak (reached at comma 20, the first block boundary, on all three models) in parentheses.

Family	Qwen3.6-27B 27B	Ouro-2.6B 2.6B	Huginn-0125 3.5B
Verbal report (top-5)	69% (AGG; rank-1 49%)	106/133 (80%; rank-1 67%)	124/126 (98%; rank-1 98%)
Introspection (8σ)	null (best rank 173 of 101)	median RR 0.50; top-1 31%	null (top-1 0%; 0/97)
Modulation topic@5 (AGG)	.38/.07/.07	.77/.39/.26	.28/.15/.19
Summoning expected@5	0.01	0.03	0.01
Two-hop probe swap	47/90 (52%); base 57/90	26/90 (29%); base 43/90	5/90 (6%); base 25/90
Flexible generalisation	80/192 (42%)	73/192 (38%); base 42/64	25/192 (13%); base 28/64
Selectivity: language (Δ@5)	+1.00	+0.25	+0.38
Selectivity: line count (AGG)	0.09/0.00	1.00/0.09	0.09/0.09
Ignition (10–90 width)	0.10	0.55	0.76
Capacity, words held	0.7 (peak 1.4)	2.4 (peak 5.0)	1.8 (peak 3.1)
Dual task, of 21 (AGG)	concept 6; maths 1	concept 17; maths 0	concept 7; maths 3

Table 6: Flexible generalisation by intervention condition — the raw data behind §6.1. Each trial swaps a source concept's lens direction for a target concept's and scores whether the answer flips to the injected concept's implication; per-category columns give flips out of 48. Qwen's run stored only the total. No-intervention baselines: Ouro-2.6B 42/64, Huginn-0125 28/64 (44%). Bottom: mean cosine between lens directions of same-category concepts (Ouro-2.6B, loop-2 band) — the category-separability comparison in Appendix M.

	Condition	Flips/192	% (95% CI)	countries	months	animals
Qwen3.6-27B	band L24–48	80	42 (35–49)	—	—	—
Ouro-2.6B	loops 2+3+4	73	38 (31–45)	33	16	24
	loop 2 only	40	21 (15–27)	31	0	9
	loop 4 only	48	25 (19–31)	30	12	6
	loops 2+3+4, 2× strength	14	7 (4–11)	10	0	4
Huginn-0125	all three windows	25	13	12	10	3
	last window only	25	13	12	10	3
	second-to-last window only	5	3	1	0	2
	all windows, 2× strength	2	1	1	0	1
Ouro-2.6B same-category direction cosine				0.36	0.67	0.39

last-window row is not merely equal in total to the all-windows row — the per-category counts are identical.

C Robustness to reasoning SFT: Ouro-2.6B-Thinking

We repeated the complete Ouro-2.6B pipeline — lens fitting on the same 1000-prompt corpus with identical hyperparameters, the structural analyses, and the full causal suite — on Ouro-2.6B-Thinking, the reasoning SFT of our base checkpoint ($\sim$8.3M supervised examples of maths, code, science and chat; the architecture is unchanged) (Zhu et al., 2025). The only pipeline change is the model identifier. One protocol note: the Thinking tokeniser emits its reasoning prefix only when explicitly requested, and under the suite's default chat template the model answers directly, so every comparison below is protocol-identical to base. All numbers in this section are re-derived by the verification suite of Appendix A.

Structure is unchanged. Deep supervision survives the SFT. The trained exit gate still concentrates its mass on the last two loops (0.37/0.46 on loops 3/4; base 0.34/0.49); logit-lens top-1 agreement at

Table 7: Direction ablation by condition — the raw data behind §6.2. Each cell is answers kept (unchanged and correct) after ablating the named direction at the intervention layers, over the prompts each model answered correctly at baseline. "Early interface only" is the loop-2 band on Ouro-2.6B and the second-to-last lens window on Huginn-0125; Qwen3.6-27B has a single interface, so the split does not apply.

Ablated direction	Qwen3.6-27B (n=36)	Ouro-2.6B (n=42)	Huginn-0125 (n=28)
Argument concept, all interfaces	26 (72%)	30 (71%)	17 (61%)
Argument concept, early interface only	—	42 (100%)	28 (100%)
Argument concept, last interface only	—	32 (76%)	17 (61%)
Answer token, all interfaces	19 (53%)	8 (19%)	0 (0%)
Unrelated word (control)	36 (100%)	42 (100%)	28 (100%)

Table 8: The causal suite on base Ouro-2.6B versus its reasoning SFT. Conventions follow Table 5; flexible-generalisation rows follow the per-loop conditions of Table 6.

Family	Ouro-2.6B 2.6B (base)	Ouro-2.6B-Thinking
Verbal report (top-5; rank-1)	80%; 67%	69%; 55%
Introspection (8σ: top-1; median RR)	31%; 0.50	31%; 0.50
Modulation topic@5, focus/control/suppress (AGG)	.77/.39/.26	.76/.45/.15
Summoning expected@5	0.03	0.03
Two-hop probe swap	26/90 (29%); base 43/90	22/90 (24%); base 35/90
Flex-gen: no-intervention baseline	42/64 (66%)	38/64 (59%)
Flex-gen: loop 2 / 3 / 4 only	21% / 29% / 25%	17% / 28% / 26%
Flex-gen: loops 2+3+4 (at 2× strength)	38% (7%)	35% (5%)
Selectivity: language (Δ@5)	+0.25	+0.00
Selectivity: line count, direct/never asked (AGG)	1.00/0.09	1.00/0.18
Ignition (10–90 width, in-band)	0.55	0.20
Capacity, words held (peak at comma 20)	2.4 (5.0)	2.1 (4.0)
Capacity, pursuit inventory ($\theta \leq 0.01$)	4.9	5.1
Dual task, of 21 (AGG)	concept 17; maths 0	concept 19; maths 0
Ablation kept: arg / answer / control	71% / 19% / 100%	66% / 26% / 97%
Ablation kept: loop-2 only / last loop only	100% / 76%	97% / 76%
Pursuit-subspace ablation kept: all / loop-2 / control	0% / 14% / 88%	0% / 11% / 87%

the four loop-end states spans 0.62–0.90 (base 0.60–0.92); and transport still factorises through the loop checkpoints, with mean composition cosine 0.962 over the 20 tested paths (base 0.971).

All interface signatures replicate. The two headline causal structure results reproduce within a few points despite the SFT: writes must still span every remaining loop (the loop-2-only clamp is half-undone by re-derivation at 17%, and loop-2-only single-direction ablation is again a near-complete null at 97% kept), and the two opposite single-loop failure modes and the 2×-strength collapse all recur. Introspection is point-identical (top-1 31%). The pursuit-based families follow suit: the inventory capacity plateau is 5.1 words against base 4.9, and the loop-2-only pursuit-subspace ablation keeps 4/38 (11%) against base 14%, with the all-interface condition at 0/38 and the random-atom control at 87%. The principal intervention-interface contrasts persist in this SFT comparison; it does not isolate an architectural cause.

Behavioural results change; intervention geometry does not. The SFT costs a few points of raw task performance under the suite's direct-answer protocol (flex-gen baseline 66%→59%, verbal report 80%→69%, probe-swap baseline 43/90→35/90), and the clamped and injected conditions move by roughly the same amount — the ratio of intervention effect to baseline is preserved.

Two notable differences. First, the language-selectivity gate collapses to always-on: the passage's language label is readable at ceiling even when task-irrelevant (Δ +0.25 → +0.00), so Ouro-2.6B's gated pattern in Table 4 does not replicate on the SFT model — consistent with reasoning SFT keeping more context features unconditionally active in the workspace. Second, ignition sharpens: the in-band 10–90 transition width drops 0.55 → 0.20 (scrambled control 0.24 → 0.12) — the Thinking model

commits more abruptly once the band resolves. Both are single-condition observations; we report them as leads rather than conclusions. Answer-direction ablation shows a smaller difference in the opposite direction: ablating the answer direction is slightly *less* destructive on Thinking (26% kept vs. 19%) while argument ablation is comparable.

D DEMONSTRATION-METRIC SEMANTICS

The metric family of §5.3 comes from the original study's interactive demonstration, whose exact semantics are not documented; we reverse-engineered them from its page bundle and reimplemented them over all three models.

Dump protocol. For each model we run the demonstration's own 8 chat prompts, generate a 128-token greedy completion, and record, at every (virtual layer, position) cell, the top-8 tokens and probabilities of both readouts: the J-lens ($\mathrm{softmax}(\mathrm{unembed}(Jh))$, through the covering lens of §3) and the logit lens ($\mathrm{softmax}(\mathrm{unembed}(h))$). These grids supply the top-8 comparison and probability-field measurements. Pursuit occupancy uses the separate reconstruction artefacts described in Appendix E.

Count (occupancy). A token's count is its number of occurrences over positions $\times$ layers $\times$ 8 slots, with whitespace-only and non-word token strings excluded. The demonstration's default layer window additionally skips the first 29% of depth; §5.3 uses full depth instead, since the depth profile is the object of study. (The skip is well chosen for a standard transformer: on Qwen3.6-27B it starts at L18, just below the content onset at L20–23 that we measure independently — but on Ouro-2.6B the same fraction leaves the content-free early layers of every loop inside the window, one of the confounds that led us to drop the demonstration's contrastive "Diff" statistic from this paper; its analysis is retained in the reproduction report.)

Probability field. The calibration measure of §5.3 is the probability a readout assigns to the token the model actually emits at the next position, averaged over positions and prompts. Because the dumps store only each cell's top 8, a token outside them contributes zero, making the reported values lower bounds; chance on the $\sim$150k vocabularies is $\sim 10^{-5}$.

Secondary occupancy prompts. The other demonstration prompts provide more specific examples. On the currency riddle ("what currency was used in the boot-shaped country before 2002?"), Ouro correctly answers "the Italian lira". The unspoken alternative *euro*, a tempting surface association between 2002 and currency, peaks in loop 2 and then declines from 259 to 188. Meanwhile, *Italy*, the intermediate concept used to derive the correct answer, increases monotonically from 177 to 390. Across all twelve grid settings, the representation therefore shifts from the rejected alternative towards the intermediate concept used by the correct answer. The same analysis is consistent with Huginn answering the riddle incorrectly ("was the Euro"): the country is scarcely recovered by the inventory (*Italy / Italian / lira / lire*: 5 slots across all 67 virtual layers, and never more than 9 anywhere on the grid), so the salient surface association (2002 + currency $\rightarrow$ euro) dominates. This is the occupancy counterpart of the two-hop weakness measured behaviourally by the probe swap (5/90, §6). Qwen's inventory also includes concepts across languages: its occupied atoms include the Chinese token for *Italy* (102 slots), alongside the English *Italy* (344) and *currency* (225). English-string token classes cannot capture this concept mass. On the sport prompt, Qwen's latent sport concepts occupy 123 band slots but none in the final eight layers, while spoken tokens retain 116 slots. The late layers therefore retain the emitted words while dropping the unspoken topic alternatives. As with the absence below L20, the zero is due to coefficient sizes (Appendix E).

Full probability-field profiles. The dumps also record, for every (layer, position), the probability the lens's decoded distribution assigns to the token the model actually emits next — $\mathrm{softmax}(\mathrm{unembed}(h))$ for the logit lens, evaluated at the true next token and averaged over positions and prompts (top-8-truncated, hence a lower bound). Chance on a $\sim$150k vocabulary is $\sim 10^{-5}$, which sets the scale. This field calibrates the *confidence* of the decodability established by rank in §5: Ouro's logit lens reaches 0.61–0.65 at *every* loop end — the checkpoint state, decoded directly, already gives the model's actual next token two-thirds probability, i.e. nearly the final layer's confidence, which

is deep supervision made visible. Huginn sits on a flat plateau of 0.02–0.045 across the recurrent stack from recurrence 3 onward (virtual layers 14–65, with a stable within-recurrence profile that peaks at each recurrence's last core layer), jumping to 0.47 only at the final coda state v66 — the state the LM head consumes. The plateau is three orders of magnitude above chance and matches the rank picture of §5.3 (top-1 hit rate 15–25% uniformly across the same depths), yet sits $20\times$ below Ouro's checkpoints — the sampled-depth model is weakly readable throughout the core, without the confidence of Ouro's checkpoints. Qwen's mid-band is near zero through the logit lens (0.003 at L36) and an order of magnitude better through the J-lens (0.019) — on a standard transformer, mid-stack states are not next-token-decodable without transport. The three-model ordering — Ouro checkpoints $\gg$ Huginn everywhere-weakly $>$ Qwen mid-band ≈ 0 — also appears in the demonstration's own probability field, without a separate set of target-token probes.

The partial-tiling reading artefact. Under the original four-lens family, later Huginn windows sit closer to their covering lens targets than earlier ones, and the resulting occupancy trend rises monotonically with depth — a curve one would naturally read as content accumulating through the recurrences. Completing the family to a full tiling — twelve additional lenses, so that a target sits at every recurrence end (v5 to v61 in steps of 4, plus the coda) and every source lies within one recurrence of its covering target — flattens the trend into the plateau of Figure 2. Under the top-8 readout, the only estimator dumped under both tilings, the partial family reads 638 total slots on the animal prompt with 32 layers at exactly zero, the full tiling 1,622 with none; under the inventory, 65 of 67 layers are occupied at the primary grid point (all 67 ungated). Within each one-recurrence window the count still rises towards the lens target — the mean-J transport attenuation of §4.1 made visible as a sawtooth — but occupancy is nonzero at every depth any lens's transport reaches. Ouro's early-loop near-zeros, by contrast, survive the within-loop distance control, but remain statements about readout recovery rather than proof of an empty state. With unevenly spaced lens targets, apparent changes across depth can be caused by changing source-to-target distance alone.

The animal prompt and replies. Figure 2 is computed over the following transcripts (the occupancy grid covers prompt and completion positions alike). Prompt, sent identically to all three models:

> I'm thinking of the largest land animal. What do they eat? Answer briefly.

Replies, verbatim:

> **Qwen3.6-27B:** "The largest land animal is the African bush elephant. They are herbivores and primarily eat grasses, leaves, bamboo, bark, roots, and fruits."
>
> **Ouro-2.6B:** "Elephants are the largest land animals and they eat a variety of plants, including leaves, bark, and fruits."
>
> **Huginn-0125:** "The largest land animal is the African elephant, which is a herbivore and feeds on a variety of plants, including grasses, leaves, and bark."

The latent/spoken classification of §5.3 follows from these texts with singular/plural variants merged: *tree* is latent for all three models, *food* for Qwen3.6-27B, *mammal* and *vegetation* for Ouro-2.6B, *fruit* and *root* for Huginn-0125 (which, uniquely, never names them).

E THE GRADIENT-PURSUIT CONCEPT INVENTORY

The occupancy estimator of §5.3 is the original study's concept inventory: a sparse non-negative reconstruction of the state from J-lens atoms. This appendix gives the algorithm, the estimator's agreement and disagreement with the demonstration's top-8 readout, and the protocols and full results of the two pursuit-based causal families (§6.2, §6.4).

Algorithm. At each (virtual layer v, position t) the dictionary is the intervention direction family of §6, unit-normalised: $a_\tau = J_v^\top u_\tau / \|J_v^\top u_\tau\|$ over the vocabulary, with whitespace-only and non-word token strings excluded exactly as in the top-8 counts (Appendix D). Greedy pursuit runs on the residual r (initialised to $h_{v,t}$): select $\arg\max_\tau \langle r, a_\tau \rangle$ among positive scores — no materialised dictionary is needed, since $\langle r, J_v^\top u_\tau \rangle$ is the lens-logit operation applied to r, normalised by precomputed atom norms — then refit all selected coefficients under a non-negativity constraint (projected gradient

on the support's Gram system, batched over positions), subtract the reconstruction, and repeat until $k_{\max} = 50$ atoms or no positive score remains. The dumps record each step's selection, a full coefficient snapshot at $k \in \{8, 16, 25, 50\}$, and the residual fraction, so every occupancy count at any (k, θ) is an offline slice of one artefact. Occupancy at (k, θ) counts the atoms among the first k whose refitted coefficient is at least $\theta\|h\|$; the body reports $k = 25$, $\theta = 0.05$, with robustness checked over the full grid $k \in \{8, 16, 25, 50\} \times \theta \in \{0, 0.01, 0.05\}$. The cells are the same (layer, position) grids as the top-8 dumps — the stored `seq_ids` reproduce each recorded forward deterministically — so the two estimators align cell-for-cell on all three models.

Agreement and disagreement with the top-8 readout. The demonstration's top-8 softmax readout — whose semantics we reverse-engineered from the study's page bundle (Appendix D) — serves as a reference estimator: every occupancy dump was also read through it. The two estimators pick genuinely different tokens: mean per-cell Jaccard overlap between the pursuit support and the top-8 slots is 0.07–0.24, and the near-twin rate (two read tokens identical up to case or plural) falls from 30–37% of cells under top-8 to 0.1–0.3% under pursuit — confirming the original study's remark that the inventory is less redundant than inner-product top-k. Despite that, both estimators recover the broad band/loop/window profiles. Individual occupancy contrasts remain sensitive to the grid setting and normalisation, as detailed below. Under top-8 the corresponding numbers are: Qwen animal prompt, 6 latent slots below L20, 1,076 latent vs. 1,116 spoken in the band, 56 vs. 231 at L55+; Ouro currency riddle, *euro* 679 $\rightarrow$ 508 against *Italy* 199 $\rightarrow$ 584 across loops; Huginn currency riddle, 2 slots for the country family over all 67 virtual layers (pursuit grid maximum: 9); Huginn full tiling, 1,622 total slots, mid-window plateau 86–108 against 21 in the first window, coda 343. Figure 3 is the top-8 counterpart of Figure 2. Two top-8 results do *not* survive the estimator change. On the animal prompt, the top-8 latent class thins fourfold against the spoken class at L55+ (the 56 vs. 231 above), whereas the inventory's latent/spoken ratio is flat-to-rising at every grid point — the thinning is a property of top-8's eight-slot competition, not of the state; the layer-level commitment that is real on Qwen lives on the sport prompt (§5.3). And Ouro's in-band excess holds under top-8 on raw loop totals, but under the inventory only as a per-layer rate (§5.3).

Variance explained. The pursuit reconstruction at $k = 50$ explains on average 4.3% of state variance on Qwen3.6-27B, 10.2% on Huginn-0125, and 21.5% on Ouro-2.6B — ordered exactly as the models' decodability (§5.3) and small in absolute terms everywhere. The k-curve has no elbow: $k = 25$ captures 91% of the $k = 50$ value on Qwen but only 78% on Ouro and 63% on Huginn, so the $k \leq 25$ cap is a convention on the recurrent models, not a saturation point. This motivates reporting occupancy with a coefficient threshold rather than raw selection counts.

The coefficient gate. Two of the §5.3 zeros — Qwen's below-L20 absence on the animal prompt and its last-8-layer sport zero — exist only under the $\theta = 0.05\,\|h\|$ gate (exact zeros at $k \geq 16$ and $k \geq 25$ respectively; near-zero at smaller k). Ungated selection fills slots at every depth: greedy pursuit always selects *something* when the atoms explain so little of the state (4.3% on Qwen), so absence registers as vanishing refitted coefficients rather than as empty slots, and the gate turns "selected at negligible weight" into "not counted". This is the role the top-8 probability floor plays implicitly for the demonstration's estimator, which shows the same below-L20 absence (6 slots) without an explicit threshold. On Qwen, absence and commitment are therefore statements about coefficient mass, not about which atoms remain selectable.

Capacity under the inventory. The capacity task of §6.4 re-scored with pursuit in place of the rank probe: identical seeded word lists and prompts, one recorded forward per trial, pursuit at the comma positions over each model's read layers. A word counts as held at a comma if, at any read layer, a selected atom with coefficient $\geq \theta\|h\|$ matches it (exact token id or case-insensitive string). Table 9 gives the last-quartile plateaus at $k = 25$.

Subspace ablation. The ten-direction ablation of §6.2 runs in two passes. An uninstrumented forward pass records, at every intervention (layer, position), the first ten pursuit atoms for that cell; each cell's atoms are orthonormalised by QR into a frozen basis Q. The intervention pass then applies $h \leftarrow h - QQ^\top h$ at those cells — removing the state's projection onto its own leading inventory subspace, as computed on the clean run. The control replaces each cell's basis with the span of ten

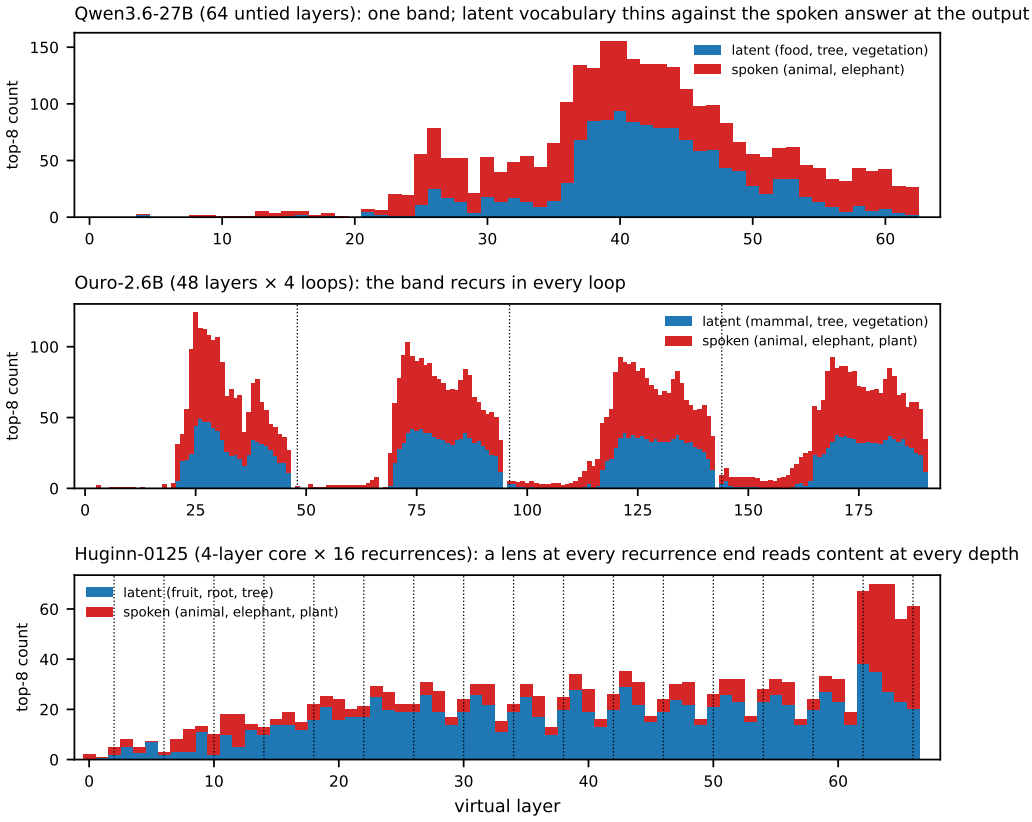

Figure 3: Count-by-layer occupancy under the demonstration's top-8 readout — the estimator-comparison counterpart of Figure 2, same prompt, same token classes, same layout. The band, loop, and window structure appears under both estimators.

Table 9: Capacity plateaus (words held, last-quartile mean) under the rank probe of Table 5 and under the pursuit inventory at $k = 25$ for three coefficient thresholds.

Model	Rank probe	$\theta = 0$	$\theta = 0.01$	$\theta = 0.05$
Qwen3.6-27B	0.7	2.0	1.9	0.1
Ouro-2.6B	2.4	4.9	4.9	4.8
Huginn-0125	1.8	3.6	3.6	2.2
Ouro-2.6B-Thinking	2.1	5.1	5.1	5.0

fixed random-vocabulary atoms (shared per layer). Trials, interfaces, and scoring are identical to the single-direction ablation family (Table 7). Table 10 gives the full counts.

Verification. All pursuit artefacts — the per-cell dumps with coefficient snapshots, the capacity dumps, and the ablation records — are released alongside the top-8 dumps, and a companion verification suite (132 checks; Appendix A) re-derives from them every pursuit number in this paper, including each robustness verdict at all twelve (k, θ) grid points.

F Worked examples

The body of the paper reports aggregates; this section shows single raw data points, so the reader can see concretely what a lens readout and an intervention record look like. Every value is pinned to its exact cell by the verification suite of Appendix A.

Table 10: Answers kept under ten-direction pursuit-subspace ablation. "Early interface only" is the loop-2 band on Ouro-2.6B and Thinking and the second-to-last lens window on Huginn-0125, as in Table 7; Qwen3.6-27B has a single interface.

Condition	Qwen3.6-27B (n=36)	Ouro-2.6B (n=42)	Huginn-0125 (n=28)	Thinking (n=38)
Pursuit subspace, all interfaces	23 (64%)	0 (0%)	1 (4%)	0 (0%)
Pursuit subspace, early interface only	—	6 (14%)	27 (96%)	4 (11%)
Pursuit subspace, last interface only	—	0 (0%)	0 (0%)	0 (0%)
Random-atom control, all interfaces	35 (97%)	37 (88%)	26 (93%)	33 (87%)

Table 11: One readout cell per model on the hidden-sport prompt: the top-5 tokens (with probabilities) of the J-lens and the logit lens at the same (virtual layer, position). "While writing" quotes the completion text around the position being read. ␣ marks a tokeniser leading space. The three cells show the three regimes the paper describes: on Qwen3.6-27B, the unspoken concept is sharp in the J-lens and invisible to the logit lens; on Ouro-2.6B, a mid-loop state read through that loop's own lens holds the candidate sports the model is still deliberating; on Huginn-0125, both lenses read the *same* word — the J-lens sharpens (0.88) what the near-token-space state already shows weakly (0.06).

Cell	Lens	Top-5 readout
Qwen3.6-27B L42, "thud of a boot connecting with a [sphere]"	J-lens	␣`football` .54, ␣`soccer` .23, ␣`Football` .09, `Football` .04, `football` .03
	logit	␣`(` .00, `...` .00, ␣`bag` .00, `sc` .00, ␣`**` .00
Ouro-2.6B loop 3, physical L26, "describe the atmosphere at a [big]"	J-lens	␣`football` .50, ␣`soccer` .21, ␣`sports` .09, ␣`rugby` .07, ␣`stadium` .03
	logit	␣`Football` .03, ␣`football` .02, ␣`championship` .01, ␣`stadium` .01, ␣`sports` .01
Huginn-0125 recurrence 15, "stadium was alive, pulsating with [the]"	J-lens	␣`excitement` .88, ␣`energy` .05, ␣`emotion` .03, ␣`enthusiasm` .01, ␣`emotions` .01
	logit	␣`excitement` .06, ␣`intensity` .04, ␣`energy` .02, ␣`anticipation` .02, ␣`electricity` .01

One prompt, three workspaces. All three readout cells come from the same demonstration prompt of §5.3, whose completion must hold a concept in mind without saying it:

> *Think of a specific sport, but don't say its name. Describe the atmosphere at a big match of that sport without naming it.*

Table 11 shows one cell of each model's (layer × position) readout grid — a high-probability cell chosen to be characteristic of that model's pattern — with the five most probable tokens under each lens.

Each cell illustrates a characteristic readout pattern for one model. Qwen3.6-27B is writing "...the heavy thud of a boot connecting with a *sphere*" — deliberately avoiding the word *ball* — and at mid-band L42 the J-lens reads the sport it is not naming at 0.54, while the logit lens at the same state reads punctuation fragments at $\sim 10^{-3}$: mid-stack content exists but is not token-decodable without transport (§5.3). Ouro-2.6B's cell sits mid-loop — between the checkpoints deep supervision trains — while its completion is still the generic "the user wants me to describe the atmosphere at a..."; the loop-3 lens reads the candidate sports the model is weighing (it later settles on rugby), where the same state's direct decode is an order of magnitude weaker. Huginn-0125's cell shows why we call its states near-token-space (§5.3): the logit lens already reads the right word everywhere, weakly, and the J-lens concentrates it.

One intervention record. The introspection experiment (§6.3) injects a concept the model is *not* computing along its lens direction, asks the model to name what it is thinking of, and records the injected concept's rank in the answer distribution at each injection strength. The per-trial records make single concepts inspectable; the concept *lightning* appears in all three models' concept lists:

Injection strength	0σ	0.5σ	1σ	2σ	4σ	8σ
Ouro-2.6B	7,524	1,809	622	114	31	**8**
Qwen3.6-27B	10,533	1,084	533	292	218	189
Huginn-0125	15,969	3,755	3,935	3,798	3,422	3,151

The verbalisation split of §6.3 is visible in this one row. As injection strength increases, the concept's rank improves monotonically in Ouro-2.6B (explicit per-iteration supervision) until it enters reportable range. The rank also improves substantially in Qwen3.6-27B but remains two orders of magnitude below that range. In Huginn-0125, the rank responds to the first increment and then plateaus: the injection measurably moves the state, but nothing carries it to the output as a word.

G ADAPTER VALIDATION AND SILENT-FAILURE TRAPS

Fitting a Jacobian lens through a recurrence exposes failure modes that do not exist on feedforward stacks. Each of the following corrupts results *without raising an error*:

1. **No-grad recurrence phases.** Huginn's iterated forward splits recurrences into no-grad and with-grad phases; passing a scalar step count places *all* of them in the former, making every Jacobian through the recurrence exactly zero. The fix is to request $(0, r)$ steps; our smoke test asserts a non-zero Jacobian at every recurrence.
2. **Stochastic state initialisation.** Huginn draws its initial recurrent state afresh per forward. The adapter seeds the generator per call (and disables test-time noise); determinism is verified bit-exact, since a lens averaged over unmatched forwards is an average over different computations.
3. **Gradient checkpointing.** Backward-pass recompute re-fires block hooks and desynchronises the virtual-layer counters; checkpointing must stay off.

The adapter's smoke tests additionally verify that each physical layer yields distinct per-firing activations, that the virtual-final state reproduces the model's own logits exactly, and that forwards are bit-exact deterministic under the seeded initialisation.

H TRANSFORMERS-5 COMPATIBILITY

Both looped models ship remote code written against transformers 4.x; our experiments run under 5.14. Ouro-2.6B needs three load-time patches: (a) `config.pad_token_id` is no longer auto-created and must be set to `None`; (b) the rotary-embedding registry lost its `'default'` (classic RoPE) entry, so a shim re-registers the default parameter function on the dynamically loaded rotary class; (c) `create_causal_mask` renamed keyword arguments. Huginn-0125 needs exactly one: it declares `_tied_weights_keys` in the 4.x list form, which crashes 5.x `post_init`; the adapter rewrites it to the dict form `{"lm_head.weight": "transformer.wte.weight"}` before `from_pretrained`. Huginn's custom recurrence cache is untested under 5.x; no experiment in this paper uses it (all forwards are cache-free and nothing calls `generate()`).

Table 12: Lens-fitting summary. Every lens uses the same 1000-prompt corpus and estimator; only the (source, target) geometry differs. The twelve Huginn gap-fill lenses put a target at every remaining recurrence end, completing the depth tiling used for readout in §5.3; the causal suite (§6) uses the four r=4/8/12/16 lenses only.

	Sources	Targets	Wall-clock
Qwen3.6-27B	all 63 layers; L0–23 (band entry)	final; L24	∼3 h; ∼2 h
Ouro-2.6B	all 191 virtual layers	virt. final + loop ends (v47/95/143)	∼1 h + 1.6 h
Huginn-0125	all 67 virtual layers	coda end + r=4/8/12; gap-fill r=1–15	∼7.3 h + per-target

I LENS-FITTING PROTOCOL AND COVERAGE

All fits use the original study's estimator on the same 1,000 WikiText-103 prompts, 128 tokens per prompt, skipping the first 16 positions. Parameters are frozen. The running-mean relative shift ends at no more than 0.016 on all Huginn lenses and approximately 5×10^{-3} on Qwen's band-entry lens. Experiments use bf16 on one to three NVIDIA B200 GPUs. The four Huginn causal lenses are the original $r = 4/8/12/16$ family; additional targets complete readout coverage but are not added to the causal suite.

J EXTENDED TRANSPORT ANALYSIS

J.1 DISTANCE AND AGGREGATE TRANSPORT

Near an Ouro target, per-prompt Jacobian norms divided by $\sqrt{d}$ are approximately 1.3–1.5, and the mean norm one block out is 1.24. Beyond one loop, the mean norm falls to 0.13–0.3. Prompt-dependent responses can cancel under averaging even though some semantic directions remain consistent. At distance 39, Qwen has norm 0.46 and top-1 agreement 2%, within Ouro's 0.14–0.53 norm range at matched distance. The Italy comparison in Table 2 is a counterexample to treating those bulk metrics as sufficient tests of semantic usability.

At Ouro's first three loop ends, direct-unembedding KL divergence from the final output is 1.7/0.5/0.1 nats, with top-1 agreement up to 92%. One block short of the final checkpoint (v190), these measures are 0.8 nats and 71%. Readability is concentrated at the states receiving explicit supervision. Across the tested loop boundary, the Italy projection retains about 0.1 of the target-state value while bulk magnitude retains 0.27–0.46. This is consistent with re-encoding of the tested semantic component; it is not a measurement of all information passing through the actual nonlinear model.

J.2 ITERATION-RELATIVE SYMMETRY

In an absolute frame, cosine between flattened Ouro final-target maps decays with virtual distance: 0.989 for adjacent sources, 0.804 at distance 48 and 0.087 at distance 190. Same-physical-layer pairs are no more similar than the all-pairs baseline (0.676 versus 0.711). This statistic does not expose a special recurrence signature towards the fixed final target.

Relative to each loop's own target, the maps are much more similar. Matched offsets on Ouro have mean cosine 0.885 (0.932 within the workspace band), versus 0.716 for mismatched offsets and 0.555 for random pairs. Top-32 singular subspaces overlap at 0.86/0.83, against a random baseline of 0.016; similarity rises to 0.988 for loops 3–4. On Huginn, independently fitted $r = 8$ and $r = 12$ lenses have matched-offset cosine 0.9991, versus 0.43 when misaligned by one recurrence (Figure 4b). These observations suggest that reusing fitted maps may reduce fitting cost, but cosine similarity alone does not establish exact interchangeability or validate transfer to every untested recurrence count.

J.3 FACTORISATION OF MEAN TRANSPORT

For each prompt the chain rule factors transport through an intermediate state. Averaging need not preserve that factorisation because $\mathbb{E}[AB]$ can differ from $\mathbb{E}[A]\mathbb{E}[B]$. Table 13 compares the direct

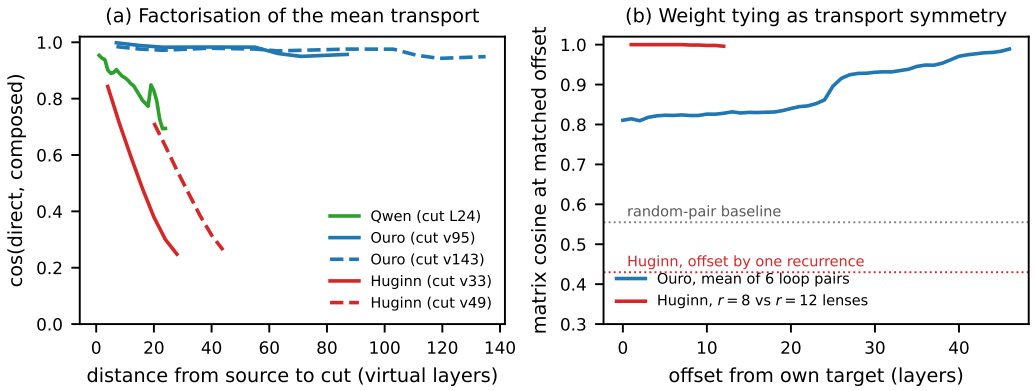

Figure 4: Transport statistics. (a) Cosine between direct mean transport and the product of mean transports through a cut, plotted against span. (b) Similarity at matched offsets from each lens's own target. Dotted baselines compare random pairs and Huginn maps misaligned by one recurrence. These are empirical matrix comparisons, not guarantees of factorisation or transfer.

Table 13: Cosine of direct and composed mean transport through a cut.

Model	Cut	Cosine
Ouro-2.6B	Supervised loop ends v95, v143	0.94–1.00
Qwen3.6-27B	Band entry L24	0.69–0.95
Huginn-0125	Recurrence ends v33, v49	0.25–0.84

mean map with the composed means. Ouro is closest to factorisation at the tested cuts; Huginn shows larger discrepancies, and Qwen lies between them.

The differences are consistent with different prompt-dependent correlations across iteration boundaries. They do not by themselves establish a Markov property, equal map magnitudes or a causal effect of deep supervision. For the measured Ouro paths, the high directional agreement motivates investigating independent per-loop fits; arbitrary long-range composition would require its own validation. Huginn's lower agreement cautions against substituting composed mean maps for end-to-end fits.

K Iteration timing and complete band maps

K.1 Ouro: checks with different measurement dependencies

Because Ouro's checkpoints are trained to be decodable (§4.1), the plain logit lens is close to a calibrated instrument *at loop ends* (0.1–1.7 nats from the final distribution, 60–92% top-1 agreement), and cross-loop timing claims can rest on it. Reading the loop ends over WikiText: loop-3's output already matches the final output for 92% of tokens; loop 4 changes the top-1 for only $\sim$8%. The model's own early-exit gate — a trained component that uses no lens at all — corroborates the same timeline, assigning exit probability $[0.03, 0.14, 0.34, 0.49]$ across the four loops.

The per-loop lens family shows which representations change. Tracking latent concepts through each loop's own band (a distance-controlled comparison, §3):

- **Easy fact** ("The capital of the country where the Eiffel Tower stands is the city of…"): *Paris* is top-1 at the end of loop 1 and every later loop end — the remaining loops maintain, not compute.
- **Riddle** ("Fact: The currency used in the country shaped like a boot is…"): loop-1's band reads only the literal surface (*boots, footwear, shoe*); **loop-2's band already reads *Italy***;

loops 3–4 maintain it. The readout identifies a correct token that is absent from the prompt, indicating that the transport retains semantic information.

- **Arithmetic** $(47 + 38)$: the answer settles in the mid-band readout only in loop 4, though the checkpoint state shows *eighty* from loop 2, suggesting that a partial answer forms early but does not settle fully until later.

Task difficulty maps onto the number of loops before the band contains the answer. Three measurements with different failure modes support this account of convergence:

1. **Per-loop J-lens readouts.** Measurement: the rank of a prompt-absent probe token (e.g. *Italy*) in $\mathrm{unembed}(J_{v\to\text{loop end}}h_v)$, at each loop's band layers. Depends on our fitted transport.
2. **Logit lens at checkpoints.** Measurement: KL and top-1 agreement between $\mathrm{unembed}(h_{\text{loop end}})$ and the model's final output distribution, over WikiText. This is the LM head applied directly to the loop-end state — exactly the operation the deep-supervision loss trained — and involves no transport.
3. **The early-exit gate.** Measurement: the model's own trained stopping signal — the learned gate that decides at each loop end whether to halt — logged per token, with no readout of any kind on our part. Its mean exit distribution over WikiText is $[0.03, 0.14, 0.34, 0.49]$ across loops 1–4: the model itself would stop at loop 3 for a third of tokens.

The first depends on the lens estimator, the second only on direct unembedding, and the third on neither — it is an internal component of the model, trained end-to-end, that we merely observe. All three agree that convergence is mostly complete by loop 3 with per-task depth, providing complementary checks with different measurement dependencies.

K.2 Huginn: prompt-dependent timing

The full heat-strip comparison is shown in Figure 5. The sushi prompt resolves early, with Tokyo readable at source steps 2–3; the Eiffel prompt yields French by step 3; the boot prompt is weaker and later. Step labels in the plots and recorded layer buckets are zero-based. These examples illustrate task-dependent timing rather than a population estimate of reasoning depth. The four lens rows read one and the same 16-recurrence computation: changing the target changes the instrument, not the generated trajectory.

L Input-injection interventions

Huginn's prelude output is re-injected at every recurrence. Its transport norm is comparatively insensitive to target distance (0.26–0.37 from prelude layer 0 and 0.54–0.77 from layer 1), while mid-recurrence sources fall below 0.06 within four to five recurrences. Ouro has a weaker analogue: the loop-1-target lens rises again near the embedding source. These structural observations motivate the following interventions on the Huginn adapter's $\mathrm{linear}(\mathrm{cat}[\text{state}, \text{embeds}])$ input. The hooks modify the injected encoding, leaving the incoming recurrent state untouched.

Attenuation. From recurrence k onwards, scale the encoding by α. At $\alpha \le 0.25$, no prompt out of eight recovers the baseline next token at any tested onset, including an onset affecting only the final four of sixteen recurrences. At $\alpha = 0.75$, seven or eight recover. With $\alpha = 0$ and $k = 8$, the latent answer's median rank is 11 through recurrence 7, rises to 2,749 at recurrence 8, and then approaches 40,000. The measured answer content therefore depends on continued refresh even after it has become readable.

Injection swap. Swap in another matched-length prompt's encoding from recurrence k onwards. The three prompt pairs differ in one word and are tested in both directions. All 30 configurations produce the new prompt's baseline answer, including $k = 12$. The old answer decays within roughly two recurrences and the new one reaches near rank 1 within roughly four. This demonstrates sensitivity to which input is refreshed, complementing the attenuation experiment. Neither intervention is a controlled comparison of models trained with and without injection; other training recipes could support different maintenance.

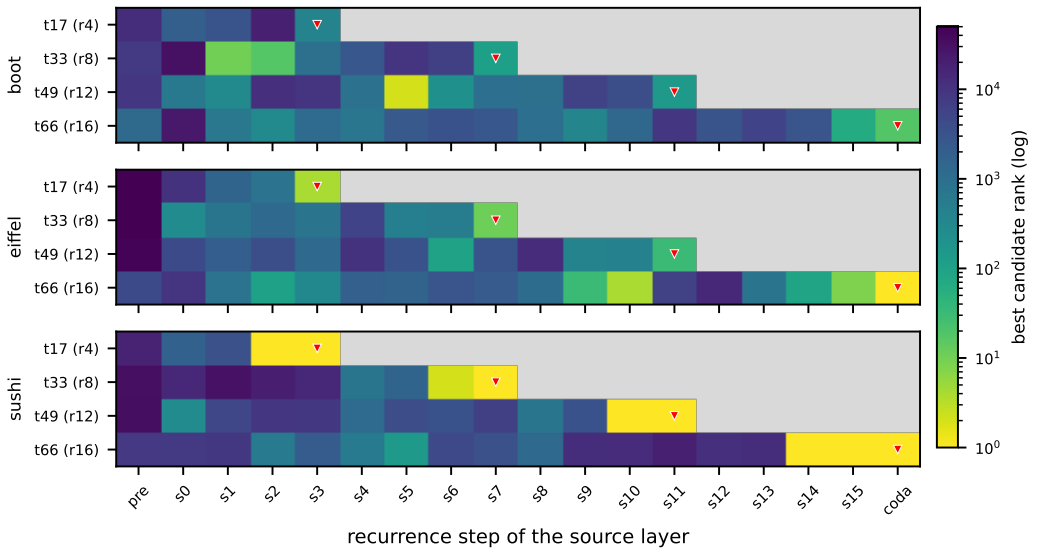

Figure 5: Full Huginn band map for the boot, Eiffel and sushi prompts. Each row is a different lens applied to the same 16-recurrence forward pass; colour gives the best candidate rank within each source-step bucket. Target triangles and grey out-of-range cells follow the main-text panel.

M Detailed causal trajectories and nulls

M.1 The two single-loop write failure modes

The loop-4-only drop (25%) cannot be re-derivation — no loop runs after it. Per-trial logit-lens rank trajectories of both answers, logged at every loop end, show the two extreme conditions fail for *opposite* reasons. Unclamped, the original answer is settled early: median rank 1 by the end of loop 2 on every category (1–3 at loop 1). Under loop-2-only clamping, failed trials show the patch *taking* and then being undone — the injected answer reaches median rank 7 at the loop-2 checkpoint but decays to 44 by loop 4 while the original recovers to rank 1: overwritten by re-derivation. Under loop-4-only clamping, failures look entirely different: the original answer already sits at rank 1 at loop-4 *entry* and the clamp lifts the injected answer only to median rank 11 — a displacement failure against consolidated content, with no re-derivation involved. Loop 3 combines the two mechanisms: it has one fewer re-derivation pass than loop 2 and less consolidated opposition than loop 4. Its result falls between the two extremes at 29% (CI 23–36), consistent with the two-mechanism account. On all-loops-clamped *successes* the original answer is driven from rank 1 to median rank 294 by loop 4 — the sustained clamp is precisely what prevents the recovery seen in the single-loop conditions. The category split matches: countries (same-category direction cosine 0.36, well-separated) succeed ∼30/48 in every condition, while months (cosine 0.67) and animals (0.39) fail in opposite directions across the two single-loop conditions (months 0 early vs. 12 late; animals 9 vs. 6, but 24 with all loops).

M.2 Task definitions and measurement limits

Language selectivity asks a question about a passage's author and tests whether the unasked language label becomes readable when task-relevant. Huginn's language hit@5 is 1.00 when relevant and 0.62 when incidental, although author accuracy is 0/8. Its no-intervention baselines are also low in two-hop probe swaps (25/90) and flexible generalisation (28/64). Intervention effects cannot be interpreted independently of these capability floors.

The capacity task presents four blocks of 20 words and probes retrieval at comma positions. Rank-probe plateaus are 0.7/2.4/1.8 words for Qwen/Ouro/Huginn, compared with approximately 25 in the original study. Pursuit at 25 atoms and a permissive coefficient threshold yields 2.0/4.9/3.6 on the

same lists (Table 9). At threshold 0.05, Qwen's plateau falls to 0.1. The sensitivity to estimator and threshold precludes interpreting these numbers as the models' total memory capacities.

Top-down summoning names a category and asks the model to bring an absent concept to mind. The score tests readability above matched foils. All three models show differences of at most 0.01 from foils, with no measured change between asking and not asking. This is a null of the manipulation/readout combination. Mathematically, any component in the nullspace of the mean Jacobian is invisible to its linear readout, whether or not it is represented elsewhere in the state. Prompt averaging and a restricted token dictionary introduce further limits. The experiments do not establish that every behaviourally idle concept is in that nullspace, or distinguish absence of storage from a failure to decode it.

N Extended related work

Global workspace accounts link access, reportability and directed modulation with recurrent processing (Baars, 1988; Dehaene & Changeux, 2011; Mashour et al., 2020). Butlin et al. (2023) discuss theory-derived indicators for AI systems; we use selected functional tests rather than drawing a consciousness conclusion. The immediate comparison is Gurnee et al. (2026), whose Jacobian-derived readouts and causal suite motivated this recurrence study.

The logit lens (nostalgebraist, 2020), tuned lens (Belrose et al., 2023) and Jacobian lens obtain their maps by direct unembedding, a learned decoding objective and estimated local transport, respectively. For weight-tied models, Lu et al. (2025) compare logit and coda lenses on Huginn and find systematic disagreement across core blocks, with little evidence for an interpretable latent chain of thought. Blayney et al. (2026) examine iteration dynamics through hidden-state geometry, attention and linear probes. Our comparison focuses on virtual-layer transport and causal workspace interfaces.

Looped transformers have also been studied as programmable computers (Giannou et al., 2023) and as learners of iterative algorithms (Yang et al., 2024). Ouro and Huginn instantiate weight reuse at language model scale (Zhu et al., 2025; Geiping et al., 2025), following earlier recurrent and parameter-sharing designs (Dehghani et al., 2019; Lan et al., 2020). The released models provide contrasting iteration interfaces, but their differing scales and training procedures prevent isolating recurrence or supervision as the cause of the observed differences.