DO MORE LOOPS FOLLOW THE EVIDENCE?

Anonymous authors
Paper under double-blind review

ABSTRACT

More inference-time computation can improve looped language models, but the same performance gain can reflect different roles for added computation. *Does another recurrent step make the model begin using facts in the prompt, or merely widen the score gap around a choice already based on them?* To answer this question, we introduce a framework that redirects the relational path supporting one candidate toward the other while keeping the question, entities, relation names, and answer pair fixed. We track which candidate wins and the score gap between candidates at successive outputs. We evaluate the framework across several recurrent systems using controlled fictional graphs and benchmark-derived multi-hop questions. Controlled fictional graphs reveal three observed profiles: acquisition, where choice control emerges at later outputs; sharpening, where control is already present and remains stable while the margin grows; and margin-control divergence, where the margin grows even as choice control declines. Repair and structural controls confirm that these effects follow the path that determines the answer rather than unrelated prompt changes. Yet control over a fixed answer pair does not guarantee reliable free-form generation. Our results show that performance gains alone cannot explain what additional recurrent computation contributes. Evaluating choice and score together reveals how evidence control changes across outputs rather than treating a larger score gap as stronger control.

1 INTRODUCTION

When a looped language model improves after another recurrent step, what has the extra computation changed? Allocating more computation at test time can improve reasoning (Snell et al., 2025). Looped language models make this idea explicit by repeatedly updating an internal state before producing an answer (Dehghani et al., 2019; Geiping et al., 2025; Zhu et al., 2025; Park et al., 2026a). Prior work shows that these additional steps can improve reasoning (Geiping et al., 2025; Zhu et al., 2025; Park et al., 2026a; Lin et al., 2026). Such gains are often summarized by plotting performance against the number of recurrent steps, which we call a depth curve. The curve shows whether later computation helps, but not what it changes. This matters for multi-hop questions: correct answers may reflect memory or partial composition, while counterfactual tests expose shortcuts that bypass prompt evidence (Biran et al., 2024; Wu et al., 2025; Yun et al., 2026). An early output may ignore the facts connecting a queried entity to an answer, then use them only after more computation. Alternatively, those facts may determine the answer from the first output while later steps widen the gap between candidate scores. The score gap can even grow as dependence on those facts weakens. A rising score-based depth curve can therefore hide different roles for recurrent computation.

We make this ambiguity observable with a paired path intervention (Figure 1). Each prompt contains two relational paths leading to fixed candidate answers. We redirect the paths so that they support the opposite candidates, then repair them to restore the original mapping, while keeping the question, entities, relation names, and answer pair fixed. If the model follows the supplied evidence, its choice should reverse after the rewire and return after repair. At each output, we record which candidate wins and the gap between the candidate scores. The **choice effect** measures whether the rewire changes the winner; the **margin effect** measures how the score gap changes. On a fictional panel that the model can solve, we call the later emergence of choice control **acquisition**. When choice control is already present at the earliest output and later computation mainly widens the score gap, we call the change **sharpening**. When the margin grows while choice control declines, we call the pattern **margin-control divergence**. Together, these measurements reveal how evidence control changes without treating a larger margin as stronger control.

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.

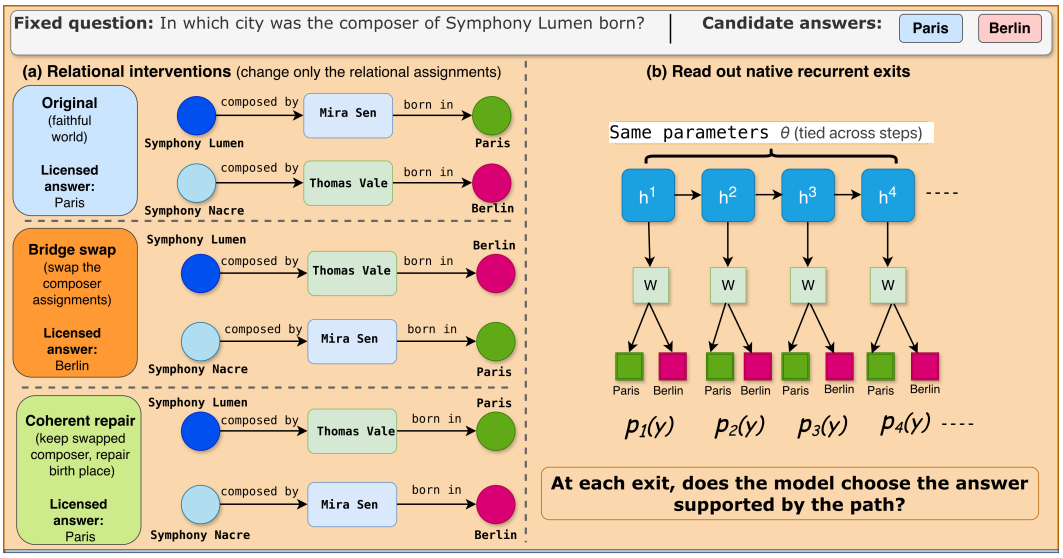

Figure 1: **A paired path intervention tests whether successive outputs follow the supplied evidence. Left:** A bridge swap changes the path-supported, or "licensed," answer from Paris to Berlin; repair restores Paris. The question and answer pair stay fixed. **Right:** A recurrent block updates the model state over repeated passes. Tracking both winner and margin reveals how evidence control changes across outputs.

We evaluate this framework on benchmark-derived questions and controlled fictional graphs across several recurrent systems (Zhu et al., 2025; Park et al., 2026a; Wang et al., 2026). The results support three conclusions. **First, recurrent systems exhibit three observed depth profiles.** At native or fixed-depth outputs, we observe acquisition when path control appears later, sharpening when early control remains stable as the margin grows, and margin-control divergence when the margin grows while choice control declines. HRM shows a related retained-to-final transition on linear, three-edge, and branching tasks. **Second, the response follows answer-bearing structure.** Repair and structural controls show that the model follows the path that determines the answer rather than a generic edit or nearby fact. **Third, training changes when control appears.** Supervising the earliest output moves control earlier. In the controlled graph-memory task, delayed control also arises with untied blocks, so parameter reuse is not necessary in that setting. Our main endpoint is a fixed answer comparison, and separate generation tests show that path control within this interface does not guarantee reliable free-form answers. Together, these results change how performance across recurrent steps should be interpreted. Evaluations should ask not only whether later computation improves an answer, but whether it changes which evidence governs the decision.

2 Experimental Design and Measurements

Setup and experimental contract. Each example asks which of two candidate answers is connected to a named source by the supplied relational facts. We score three versions of the same example. The original graph supports one candidate, the bridge swap redirects the path to the other candidate, and the repair restores the original source-to-answer mapping. Across these versions, the question, entities, relation names, and answer pair remain fixed. At every *readout*, meaning each evaluated stage at which a model state is mapped to candidate scores, we record which candidate wins and the gap between the candidate scores. A model that follows the supplied evidence should reverse its choice after the swap and return after repair.

Paired path intervention. Each unit contains two relational chains, $S_1 \to B_1 \to Y_1$ and $S_2 \to B_2 \to Y_2$. Here, S is the source named in the question, B is an intermediate bridge, and Y is a candidate answer. The original graph connects S_1 to Y_1 and S_2 to Y_2. The bridge swap exchanges the first links, causing each source to reach the other answer. Repair leaves the first links crossed but exchanges the

final links, restoring the original source-to-answer mapping. The question, entities, relation names, answer pair, record order, and candidate order otherwise remain fixed. We query both sources and use both candidate orders, ruling out fixed rules based only on the source, answer identity, or answer position. The three-edge task inserts another intermediate node, $S \to B \to M \to Y$; the swap changes the first link and repair changes the last.

Candidate scores. On natural questions, the *margin* is the difference between the candidates' complete-answer log probabilities under teacher forcing, where each token is conditioned on the prompt and preceding answer tokens. The *first-divergence choice* instead compares token scores, or logits, at the first position where the candidates differ, so it need not match the sign of the complete-answer margin. Ouro's scored continuation includes the space before the answer; LoopUS scores the bare candidate at the assistant boundary. Section 5.1 repeats Ouro's comparison without the space. On fictional tasks, A and B are single-token answers, so choice and margin use the same logit difference. Both measurements compare predefined candidates rather than test free-form generation, which we evaluate separately in Section 5.

Path effects at one readout. For each query and candidate order, the answer supported by the original graph is the reference. At readout k, let m_k^{orig} be the reference answer's score minus the other answer's score on the original prompt. Let m_k^{swap} be the same signed margin after the bridge swap. For graph state s, let c_k^s equal 1 when the reference answer wins, 0 when the other answer wins, and $1/2$ for a tie. Averaging over paired examples gives

$$D_k = \mathbb{E}\Big[m_k^{\text{orig}} - m_k^{\text{swap}}\Big], \qquad C_k = \mathbb{E}\Big[c_k^{\text{orig}} - c_k^{\text{swap}}\Big]. \tag{1}$$

Here, D_k is the **raw-margin effect** and measures how strongly the rewire changes the score gap. C_k is the **choice effect** and measures how strongly the rewire changes the winner. Positive values mean that the readout changes in the direction predicted by the supplied path. Multiplying every score difference by the same positive factor can increase D_k, but it cannot change a winner and therefore leaves C_k unchanged.

Change across readouts. For an earlier readout a and a later readout b, we define $\Delta D_{b,a} = D_b - D_a$ and $\Delta C_{b,a} = C_b - C_a$. A positive ΔC is a *choice gain*. On fictional panels, we call the transition **acquisition** only when the model passes the competence gate and the study-specific decision rules. On benchmark-derived questions, we report the same quantity as a secondary choice gain rather than label it acquisition. When choice control is already present, the interval for $\Delta D_{b,a}$ is positive, and the interval for $\Delta C_{b,a}$ includes zero, we call the change **sharpening**. When the intervals instead support $\Delta D_{b,a} > 0$ and $\Delta C_{b,a} < 0$, we call the pattern **margin-control divergence**.

Repair and uncertainty. Repair recovery is the pooled mean repair-minus-swap effect divided by the pooled mean original-minus-swap effect, using margins (R^m) or choices (R^c). One indicates full recovery and values above one indicate overshoot. The branching task applies the same ratio to its relation interaction. Natural margins are log probabilities and fictional margins are logits, so we do not compare their magnitudes. All intervals use 10,000 paired bootstrap samples over independent units, resampling each unit's graph states, candidate orders, and readouts together. Appendices F–G provide the exact scoring conventions, decision rules, and resampling schemes.

Models and readout interfaces. We call an output *native* when it is produced by the released model's ordinary forward pass and output head. A *trained exit* is an output that was explicitly supervised or designated during training. These properties are not equivalent. For Ouro, LoopUS, and Nanbeige, K counts recurrent applications: K1 is the readout after one application, K2 after two, and so on. HRM uses a different hierarchy. H1 is the state after its first slow recurrent cycle and H2 is the state after its second; each slow cycle contains three faster inner cycles. These labels describe progression within a model, not equal amounts of computation across models. We therefore compare readouts only within the same system.

Native exits: Ouro-2.6B produces trained outputs after four successive recurrent passes in one run, denoted K1–K4 (Zhu et al., 2025; ByteDance, 2025). **Separate fixed-depth outputs:** each LoopUS point comes from a separate execution stopped after the requested number of recurrent applications rather than from one shared trajectory. The natural panels use an author-released 1.7B checkpoint, the first fictional panel uses an 8B checkpoint, and an additional confirmation evaluates the released Phi-4 LoopUS checkpoint at K1, K2, K4, and K8 (Park et al., 2026a;b;c;d). Nanbeige4.2-3B exposes its loop count through the released model interface. We execute its shared 22-layer stack once or

twice and apply the released output head, producing native one-pass and two-pass outputs. Only the two-pass output is the checkpoint's trained and default endpoint, so K1 is a direct model output but not a trained intermediate exit (Nanbeige, 2026).

Intermediate-state readouts: HRM-Text-1B normally returns H2. We retain H1 from the same forward pass and apply the released H2-trained output layer to it. Separate linear classifiers at H1 and H2 test whether the graph-supported answer can be recovered on unseen worlds and relation families (Wang et al., 2026; Sapient Intelligence, 2026a). These classifiers are analysis tools rather than model outputs. For standard Qwen3-4B, we retain the state after every Transformer block and apply the released final normalization and output head. We label these projections L1–L36 by block number. L1–L35 are fixed-head intermediate-layer readouts; only L36 is the model's native output (Qwen Team, 2025).

K, H, and L therefore refer to different computations. We verified checkpoint revisions, configuration fields, prompt templates, and executed block counts against the released implementations. The LoopUS findings concern the evaluated author releases; we did not reconstruct their training recipes. Appendix B records the architectures, checkpoints, and execution checks.

Benchmark-derived panels. These panels retain natural questions and answer identities while replacing the original passages with the four relational statements needed for the two chains. Each matched unit pairs two answerable chains and evaluates both questions, all three graph states, and both candidate orders. The Ouro panels contain 124 2WikiMultiHopQA pairs and 95 MuSiQue pairs. The LoopUS panels contain 93 and 95 pairs, respectively (Ho et al., 2020; Trivedi et al., 2022).

Fictional panels. The main panel contains 288 balanced worlds with opaque entities, identifiers that never recur across worlds, two alternatives represented by stable single tokens, and six families of fictional relation names. Most models use A and B; Nanbeige uses two displayed terminal labels. A rule-based solver verifies every graph state before model evaluation. Separate panels contain 144 HRM branching worlds and 144 three-edge worlds for each of Ouro and HRM.

Competence before intervention. Every intervention study first checks accuracy on the unchanged task. Ouro and LoopUS-8B use all 288 fictional worlds after a 48-world original-only competence check. HRM, Nanbeige, and Phi-4 use those 48 worlds only for competence and the remaining 240 worlds for intervention. On fictional tasks, accuracy must reach 0.65, its 95% lower confidence bound must exceed 0.50, and accuracy must reach 0.55 for every query arm and candidate order. Only then do we evaluate swaps and repairs. HRM and the three-edge studies use separate competence and intervention worlds. The original Ouro natural panels report accuracy descriptively, while the LoopUS natural panels use the task-specific checks in the appendix. Demonstrations and prompt formats are selected on development worlds and then held fixed. Appendices D–E provide the panel construction, prompts, and data splits.

3 THREE OBSERVED DEPTH PROFILES OF EVIDENCE CONTROL

Controlled graphs reveal three observed depth profiles. Ouro and Nanbeige show acquisition at later outputs, while HRM shows a related transition from a retained state to its final output. LoopUS-8B shows sharpening: choice control is already strong and the margin grows. Phi-4 LoopUS shows margin-control divergence: its margin grows while choice control declines. A wider margin can therefore accompany increasing, stable, or declining choice control. Figure 2 illustrates acquisition and sharpening, while Table 1 summarizes all three profiles. These comparisons span trained exits, native and separate fixed-depth outputs, and retained-state readouts, so all depth comparisons are within-system.

Acquisition: control appears at later readouts. Unique opaque entity names prevent reuse of memorized facts about those entities. On 288 two-edge worlds, Ouro's choice effect rises from 0 at K1 to 0.480 at K4, and repair restores the K4 margin effect. A retrospective analysis of these saved outputs shows that positive choice control rarely reverses. The K4-to-K1 gain also remains positive in every K1-margin quartile and across four prompt formats. Nanbeige shows a related transition through a different interface: on the 240 worlds held out from its competence screen, its choice effect rises from 0.020 after one pass to 0.597 after two, a gain of 0.578, and repair restores the removed margin. HRM's choice effect increases by 0.896 from retained H1 to native H2 when both are read through the released H2-trained head. Separately fitted linear classifiers preserve this gap on held-out

relation families, with probe-based choice effects of -0.004 at H1 and 0.989 at H2. This weakens a simple linear-head mismatch explanation, but these classifiers measure decodability rather than native choices, and nonlinear or distributed information may still be present at H1. Appendix I gives the complete controlled-graph trajectories, retrospective checks, and prompt-format results.

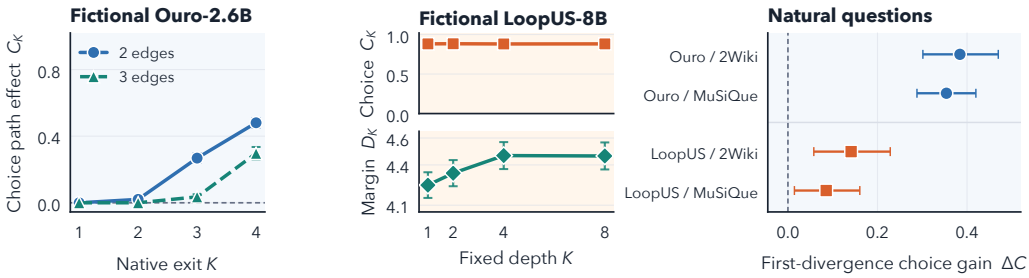

Figure 2: **Representative results show acquisition and sharpening. Left:** Ouro's choice effect grows from near zero across separate two- and three-edge fictional panels. **Middle:** LoopUS-8B already has strong control at K1; its choice effect stays flat through K8 while the raw-margin effect grows. **Right:** Secondary analyses on natural questions show positive first-divergence choice gains. Whiskers show 95% bootstrap intervals; Table 1 includes Nanbeige, Phi-4, and HRM.

Table 1: **Choice and margin reveal three observed depth profiles.** Entries are later-minus-earlier contrasts with 95% bootstrap intervals; the branching row is relation-conditioned. Repair uses margin (R^m) or choice (R^c); one denotes full average recovery. Comparisons are within-system, and margin scales are not compared across systems.

System and task	Readouts	Margin change	Choice change	Repair recovery
Acquisition and related later-control gains on fictional two-edge graphs				
Ouro-2.6B 2 edges	K4–K1	0.277 [0.264, 0.290]	0.480 [0.451, 0.509]	$R^m = 1.032$ [1.005, 1.058]
Nanbeige4.2-3B 2 edges	K2–K1	6.041 [5.781, 6.316]	0.578 [0.544, 0.611]	$R^m = 1.114$ [1.066, 1.164]
HRM-Text-1B 2 edges	H2–H1	3.768 [3.622, 3.915]	0.896 [0.874, 0.916]	$R^m = 1.045$ [1.027, 1.062]
Sharpening profile: early control and later margin growth				
LoopUS-8B 2 edges	K8–K1	0.217 [0.208, 0.225]	0.000 [−0.010, 0.010]	$R^m = 1.082$ [1.061, 1.104]
Margin-control divergence: margin grows as choice control declines				
Phi-4 LoopUS 2 edges	K8–K1	**0.035** [0.029, 0.041]	−0.062 [−0.089, −0.036]	$R^m = 1.023$ [1.005, 1.042]
Fictional task extensions				
Ouro-2.6B 3 edges	K4–K1	0.348 [0.323, 0.374]	0.297 [0.260, 0.335]	$R^c = 0.892$ [0.776, 1.012]
HRM-Text-1B 3 edges	H2–H1	1.924 [1.785, 2.065]	0.491 [0.444, 0.537]	$R^c = 1.014$ [0.930, 1.101]
HRM-Text-1B branching	H2–H1	0.867 [0.816, 0.917]	0.562 [0.526, 0.596]	$R^c = 1.009$ [0.966, 1.052]
Benchmark-derived questions: margin in nats; first-divergence choice is secondary				
Ouro-2.6B 2Wiki-based	K3–K1	3.115 [2.722, 3.532]	0.384 [0.301, 0.470]	$R^m = 0.711$ [0.623, 0.784]
Ouro-2.6B MuSiQue-based	K3–K1	3.141 [2.776, 3.508]	0.354 [0.288, 0.420]	$R^m = 0.648$ [0.561, 0.726]
LoopUS-1.7B 2Wiki-based	K8–K1	1.108 [0.835, 1.390]	0.141 [0.058, 0.228]	$R^m = 0.790$ [0.710, 0.859]
LoopUS-1.7B MuSiQue-based	K8–K1	0.687 [0.483, 0.915]	0.086 [0.014, 0.161]	$R^m = 0.670$ [0.573, 0.753]

Sharpening: early control and later margin growth. On the same two-edge task, LoopUS-8B has a choice effect of 0.882 at both K1 and K8, while its raw-margin effect increases by 0.217. To test whether the larger margin merely reflects a common rescaling of scores, we fit one positive multiplier on five relation families and evaluate it on the sixth. After rescaling, the held-out difference between K1 and K8 is essentially zero, consistent with score scaling rather than stronger choice control.

Margin-control divergence: margin growth can accompany weaker control. A separate Phi-4 LoopUS checkpoint begins with strong K1 control, but its K8-to-K1 choice change is $-0.062\,[-0.089, -0.036]$ while its raw-margin change is $0.035\,[0.029, 0.041]$. Repair restores the K8 margin effect. Within this assay, margin growth is therefore not sufficient evidence of stronger choice control. These fixed-depth aggregates do not explain why the measures move in opposite directions. LoopUS-1.7B's positive natural-question gains show that the observed profile depends on the checkpoint, task, and interface.

Beyond the original two-edge task. On separate 144-world three-edge tests, Ouro shows acquisition across its trained K1–K4 exits, while HRM shows a corresponding retained-to-final transition from H1 to H2. Both produce positive choice gains and near-full repair. HRM also shows a retained-to-final control gain on a branching task, where the relation named in the question selects one of two outgoing links: the relation-conditioned choice effect is 0 at H1 and 0.562 at H2. These results broaden the finding across path structures, but they do not imply one loop per edge. Appendix J reports the complete three-edge experiments.

Benchmark-derived questions. Complete-answer margins grow across depth for Ouro and LoopUS on both benchmark-derived panels, and coherent repair restores 64.8–79.0% of the later margin effect removed by the swap. At K3, Ouro's pooled graph-consistent two-candidate accuracy is 0.690 on 2Wiki and 0.675 on MuSiQue. The secondary first-divergence analysis also shows later choice gains: 0.354–0.384 across Ouro's trained exits and 0.086–0.141 across LoopUS-1.7B's fixed-depth runs, mostly by K2. A positive rescaling cannot change the winning candidate, so these gains indicate greater sensitivity to the supplied path at the choice level. We report them as choice gains rather than assigning an acquisition label. Ouro's scores include the space before each answer, while LoopUS uses bare candidates at the assistant boundary. Section 5.1 examines Ouro's comparison without the space. Appendix H gives the complete trajectories and accuracy results.

Section 4 next asks whether this sensitivity follows answer-bearing connectivity rather than merely any prompt edit.

4 THE EFFECT FOLLOWS THE PATH, NOT JUST THE EDIT

Section 3 showed that a bridge swap changes which candidate wins more strongly at later readouts. We now test two simpler explanations: that the model reacts to any edited fact, or that it follows the final fact placed nearest the swapped bridge in the prompt. On 144 held-out fictional worlds for Ouro, we first apply the same edit either on or off the queried answer path. We then make graph connectivity and prompt proximity support different answers.

The same edit matters only on the queried path. Within each fictional world, the evidence text and edited positions remain identical. Only the question and associated answer pair change, determining whether the swapped component lies on the answer path. The on-path and off-path choice effects are both zero at K1. By K4, the on-path effect reaches 0.482, while the off-path effect remains near zero at 0.018. The corresponding margin effects are 0.270 and approximately zero. The edit's presence, wording, and position therefore cannot by themselves explain the on-minus-off separation.

Connectivity survives conflicting proximity. To test a nearest-record explanation, we keep the swapped first-step facts fixed and vary which final fact completes the path and where it appears. In the matched condition, connectivity and the nearest final fact support the same answer. In the crossed condition, the nearest fact supports the competing answer. The order of demonstrations and the position of relevant information can affect Transformer predictions (Lu et al., 2022; Liu et al., 2024; Ok & Lee, 2026). At K2, the choice effects are 0.204 when connectivity and proximity agree and 0.007 when they conflict. By K4, they reach 0.740 and 0.438. At K4, both the on-minus-off difference and the crossed graph effect have positive point estimates in all six relation families. Proximity therefore changes when the response appears and how strong it becomes, but the aggregate choice effect still moves toward the answer selected by connectivity.

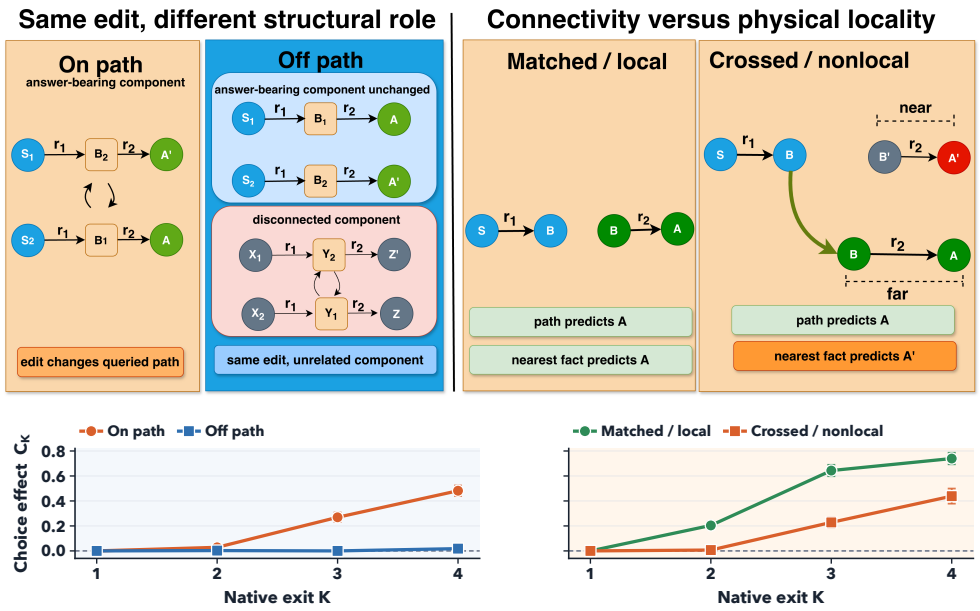

Figure 3: **The response tracks connectivity; proximity changes its timing and strength. Top:** On-path/off-path and matched/crossed controls separate answer-bearing connectivity from edit presence and final-fact proximity. **Bottom:** Choice effects across Ouro's native exits on 144 test worlds; whiskers show 95% bootstrap intervals.

Natural questions reproduce the distinction. Here, on-minus-off subtracts the response to the identical swap in a disconnected component from its response on the queried path. At Ouro K3 and LoopUS K8, the complete-answer on-minus-off effects are 5.120 and 3.510 nats. Ouro's corresponding first-divergence effect rises from 0.002 at K1 to 0.516 at K3. Under adverse proximity, its graph-consistent choice effect rises from 0.007 to 0.211. LoopUS already shows on-minus-off choice control at K1, increasing from 0.316 to 0.394 at K8. These within-model categorical estimates use leading-space candidates for Ouro and bare tokens at the assistant boundary for LoopUS; Section 5.1 audits Ouro's boundary sensitivity. Appendix K gives the complete designs, trajectories, and intervals.

These controls isolate answer-bearing connectivity from edit presence and proximity, but do not locate composition or establish a graph algorithm. Section 5 asks when control becomes readable and whether it extends to generation.

5 What shapes when control becomes readable?

Released checkpoints differ in architecture, size, data, and training, so their trajectories cannot explain what determines timing. We therefore train small Transformers on a symbolic graph-memory task while varying parameter sharing and supervision. One model reuses the same block four times; the other uses four untied blocks. Supervision begins at K4, K3, K2, or K1 and continues through the final readout. Every condition reaches at least 99.8% accuracy at K4.

Supervision moves control earlier. K1 choice control remains near zero when supervision begins at K2 or later. When all readouts are supervised, it reaches 0.567 for tied blocks and 0.827 for untied blocks (Figure 4a). All conditions end with nearly complete K4 control, so supervision changes when control becomes readable without materially changing final competence. Because the total loss weight is fixed, adding K1 also reduces the weight assigned to later readouts; the experiment does not separate these two changes.

Weight sharing is not required. With K4-only supervision, tied and untied models both move from near-zero K1 control to nearly complete K2 control. Delayed control can therefore arise without

parameter reuse on this task, although the experiment does not establish that tied and untied models learn the same representations.

Readout analyses preserve the delayed transition. Separate linear classifiers remain at chance at K1 and become nearly perfect at K2, weakening a simple mismatch with the K4-trained head while leaving open nonlinear information at K1. Qwen3-4B shows a related fixed-head transition across 36 untied blocks on 288 fictional worlds (Figure 4b): layer 24 is the first evaluated layer whose interval excludes zero, and the effect reaches 0.582 at the native final output. The trajectory is nonmonotonic. Only the final layer was trained for direct decoding through this head, so earlier projections may miss available information. This is a result about standard Transformer depth, not recurrent inference-time passes. Appendix L gives the complete training and readout analyses.

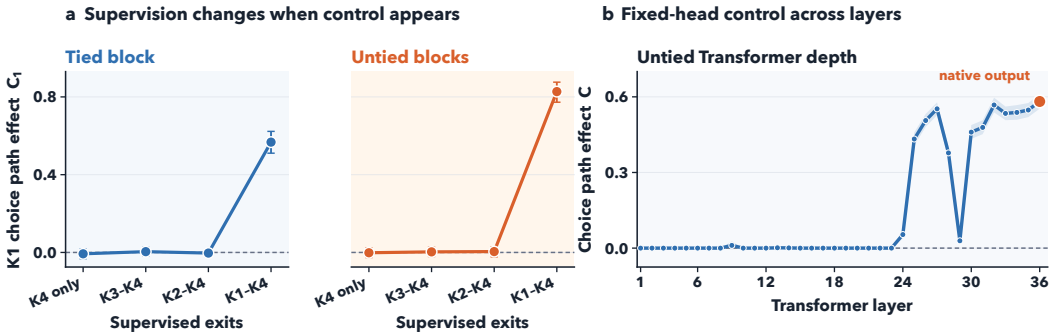

Figure 4: **Supervision changes when control appears; a standard Transformer shows a related fixed-head transition. (a)** K1 control remains near zero unless K1 is supervised, for both tied and untied blocks. **(b)** Qwen3-4B's final head reveals a late, nonmonotonic transition across 36 blocks; orange marks the native final output, while earlier points are fixed-head layer readouts. Error bars and shading show 95% bootstrap intervals.

5.1 Path control does not guarantee reliable generation

The fixed-choice results do not imply that a model will generate the graph-supported answer without a supplied comparison. We test this boundary by removing the displayed answer list, evaluating free-form outputs, and rescoring candidates without the leading space used in Ouro's primary analysis. Appendix M gives the generation metric, parser, and complete intervals.

Most of the effect remains without displayed choices. Removing the candidate list from Ouro's 2Wiki prompts retains 88.5% of the K3 complete-answer margin effect and 90.1% of its K1-to-K3 gain. This does not establish equivalence: the answer strings remain in the evidence, and the evaluation still scores predefined continuations. It does show that copying from the displayed candidate list cannot explain the effect.

Directional sensitivity survives, but reliable generation does not. Without an answer list or supplied continuation, rewiring shifts Ouro's outputs toward the graph-supported answer on both natural panels. Whole-output accuracy is only 0.540 on 2Wiki and 0.503 on MuSiQue, with both intervals including 0.50. None of the tested Ouro or LoopUS generation settings passes its predefined criterion.

The token boundary changes individual winners. Removing the leading space reverses the preferred answer on 34.4% of 2Wiki rows and 35.4% of MuSiQue rows for which both scoring rules select a candidate. The K1-to-K3 choice gain nevertheless remains positive under bare-token scoring on both datasets. The interface therefore changes many individual decisions without erasing the aggregate depth effect.

6 Related work

Recurrent and adaptive computation. Prior work develops recurrent, adaptive-depth, latent-refinement, hybrid, training-free, and slow–fast systems (Dehghani et al., 2019; Graves, 2016; Banino et al., 2021; Schuster et al., 2022; Geiping et al., 2025; Zhu et al., 2025; Park et al., 2026a; Fu et al., 2026; Wang et al., 2026; Lee et al., 2026; Chen et al., 2026; Saunshi et al., 2025; Kohli et al., 2026). These systems show that repeated computation can help; we ask whether it changes which supplied evidence controls the answer. **Intermediate readouts and interventions.** Prior work traces readability and uses mediation, activation interventions, and representation inspection to locate computation (Lam-Muir, 2026; Curth et al., 2026; Popescu et al., 2026; Lin et al., 2026; Wang & Reid, 2026; Vig et al., 2020; Geiger et al., 2021; Meng et al., 2022; Conmy et al., 2023; Ghandeharioun et al., 2024). We instead change which answer the prompt evidence supports and measure its effect on choice. **Multi-hop reasoning and path use.** Benchmarks and analyses of multi-hop reasoning study composition, shortcuts, memorization, and distractors (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022; Press et al., 2023; Biran et al., 2024; Yang et al., 2024; 2025; Wu et al., 2025; Yun et al., 2026). Mechanistic and behavioral studies examine entity binding, factual recall, context position, and answer presentation (Feng & Steinhardt, 2024; Geva et al., 2023; Liu et al., 2024; Wong et al., 2026). We hold the question, entities, relations, and answer pair fixed, rewire path connections, and track choice and margin.

7 Discussion

Additional computation can play different roles. Depth curves hide three distinct changes: acquisition, where path control emerges; sharpening, where established control remains stable while the margin grows; and margin-control divergence, where the margin grows while choice control declines. Phi-4 directly refutes reading a wider margin as stronger evidence use: its raw-margin effect rises as its choice effect falls. A recurrent step should therefore be characterized by what changes in the decision, not only whether an endpoint improves. What widening scores amplify when control weakens remains open.

The timing of control is not the timing of representation. Repair, off-path edits, and conflicting-proximity controls establish that answer-bearing connectivity causally influences the predefined choice. They do not show when the relevant facts first become represented, where the path is composed, or whether the model implements a graph algorithm. Information being present, readable through a chosen head, and controlling a decision are different claims; our intervention addresses the third. The HRM probes test linear readability but cannot exclude nonlinear or distributed information at H1. The controlled training study shows that supervision can move control earlier without changing final competence, while terminal-only tied and untied models can both show delayed control. Timing can therefore reflect training pressure as well as task computation, although loss-weight redistribution and the small symbolic task limit that conclusion.

Evidence control depends on the answer interface. Removing the answer list preserves most of the likelihood effect, yet no generation setting passes its competence criterion, and token boundaries change many winners. Evidence can therefore govern a supplied candidate comparison without reliably governing free-form generation. This interface gap is itself informative: representation, decision control, and generation place different demands on a model. Evaluations should establish competence, identify whether each readout is native, trained, or analytical, measure choice alongside margin, use repair and structural controls for evidence-use claims, and test generation separately. **Recurrent models should be evaluated by when supplied evidence controls a decision, what later computation amplifies, and whether control survives into generation.**

8 Conclusion

More loops do not have a single effect. Swap-and-repair reveals acquisition, sharpening, and margin-control divergence: control can grow, remain stable, or decline while margins grow. Structural controls tie these changes to answer-bearing connectivity, yet fixed-choice control does not guarantee reliable generation. Depth curves are therefore incomplete: evaluation must ask whether additional computation changes **which evidence governs the decision**.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing and polishing, including improving grammar, clarity, readability, and presentation of author-written text. We also used generative AI tools to assist with code generation and editing for experimental workflows and to identify a broad set of potentially relevant literature. All AI-assisted code was reviewed and validated by the authors, and all candidate literature was independently reviewed and selected by the authors. Generative AI tools were not used to generate or refine research hypotheses, interpret experimental results, or formulate scientific conclusions. All AI-assisted outputs were reviewed by the authors, who take responsibility for the final content of this work.

ETHICS STATEMENT

This study uses generated fictional graphs and public or access-controlled research datasets and model checkpoints under their stated terms. It involves no human participants, private user data, or deployment. The behavioral results do not establish human-like reasoning or suitability for high-stakes use.

REPRODUCIBILITY STATEMENT

Appendices A through N specify the checkpoints, panel construction, model-facing prompts, execution checks, measure definitions, and resampling procedures used for the central results. The accompanying anonymous supplementary artifact contains the generated panels, evaluation and analysis code, saved row scores, and commands for checking release integrity and reproducing the reported tables, figures, measures, and 95% confidence intervals on CPU. Model checkpoints and licensed benchmark text are not redistributed; pinned revisions, access requirements, and optional model re-execution instructions are documented in the artifact.

REFERENCES

Andrea Banino, Jan Balaguer, and Charles Blundell. PonderNet: Learning to ponder. In *8th ICML Workshop on Automated Machine Learning*, 2021. URL https://openreview.net/forum?id=1EuxRTe0WN.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14113–14130. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.781. URL https://aclanthology.org/2024.emnlp-main.781/.

ByteDance. Ouro-2.6B. Hugging Face model card, 2025. Model card; experiments use revision 1ed04250da1a9936042725d302e81c8fa2ab5abd.

Lizhang Chen, Jonathan Li, Chen Liang, Ni Lao, and Qiang Liu. Training-free looped transformers, 2026. URL https://arxiv.org/abs/2605.23872.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352. Curran Associates, Inc., 2023. doi: 10.52202/075280-0719. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Alicia Curth, Rachel Lawrence, Sushrut Karmalkar, and Niranjani Prasad. Do transformers use their depth adaptively? evidence from a relational reasoning task, 2026. URL https://arxiv.org/abs/2604.12426. ICLR 2026 Workshop on Logical Reasoning of Large Language Models.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

DiscoLoop. Disco-7T. Hugging Face repository, 2026. Configuration inspected at revision ad432ee2685ef78db551614c3ebc9f724f627a57.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *International Conference on Learning Representations*, volume 2024, pp. 36391–36413, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9d1b7fc578c0d2d6431fc26d736ecaf3-Abstract-Conference.html.

Hengyu Fu, Tianyu Guo, Zixuan Wang, Hanlin Zhu, Jason D. Lee, Jiantao Jiao, Stuart Russell, and Song Mei. DiscoLoop: Looping discrete embeddings and continuous hidden states for multi-hop reasoning, 2026. URL https://arxiv.org/abs/2607.00341.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. 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, pp. 41340–41391. Curran Associates, Inc., 2025. doi: 10.52202/085713-1380. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3b01972cf31e6fa0fe29e4b8b5c2a0a1-Abstract-Conference.html.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15466–15490. PMLR, 2024. URL https://proceedings.mlr.press/v235/ghandeharioun24a.html.

Alex Graves. Adaptive computation time for recurrent neural networks, 2016. URL https://arxiv.org/abs/1603.08983.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Harsh Kohli, Srinivasan Parthasarathy, Huan Sun, and Yuekun Yao. Loop, think, & generalize: Implicit reasoning in recurrent-depth transformers, 2026. URL https://arxiv.org/abs/2604.07822. Accepted at the Conference on Language Modeling (COLM) 2026.

Simon Lam-Muir. The ignition is real, and it lives at the readout: Latent composition, difficulty-clocked ignition, and the interface-constituted commit in a recurrent-depth reasoner, 2026. URL https://arxiv.org/abs/2608.03263.

Ryan Lee. anchorfix-base-1e18-d896-seed42. Hugging Face model card, 2026a. Model card and configuration; revision 95e56a838fb7af0d2d13a6078a2b49c280dc36e3.

Ryan Lee. anchorfix-looped-1e18-d896-seed42. Hugging Face model card, 2026b. Model card and configuration; revision 6be97db079ff30d4e1d465512d55aaefe1b24395.

Ryan Lee, Jacob Biloki, Edward J. Hu, and Jonathan May. Sparse layers are critical to scaling looped language models, 2026. URL https://arxiv.org/abs/2605.09165.

Ruhai Lin, Yiyang Guo, Rui-Jie Zhu, Hao Ye, and Jason K. Eshraghian. Allocating recurrent compute in looped language models, 2026. URL https://arxiv.org/abs/2608.18230.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Nanbeige. Nanbeige4.2-3B. Hugging Face model card, 2026. Model card; experiments use revision b82e54bd609793562a75cbf9337970a93369eab5.

Hyunjong Ok and Jaeho Lee. Lost in the prompt order: Revealing the limitations of causal attention in language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38566–38587. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1921. URL https://aclanthology.org/2026.findings-acl.1921/.

Taekhyun Park, Yongjae Lee, Dohee Kim, and Hyerim Bae. LoopUS: Recasting pretrained LLMs into looped latent refinement models, 2026a. URL https://arxiv.org/abs/2605.11011.

Taekhyun Park, Yongjae Lee, Dohee Kim, and Hyerim Bae. Qwen3_1.7B_LoopUS_SFT. Hugging Face model card, 2026b. Model card; experiments use revision 469f03291c3be06d7a428a4043809d818c304a25.

Taekhyun Park, Yongjae Lee, Dohee Kim, and Hyerim Bae. Qwen3-8B_LoopUS. Hugging Face model card, 2026c. Model card; experiments use revision 4424cfd8e36c77fee337a04602e3ddf6faae533f.

Taekhyun Park, Yongjae Lee, Dohee Kim, and Hyerim Bae. Phi4_LoopUS. Hugging Face repository, 2026d. Repository snapshot and configuration; experiments use revision c8e514aef96aa7106fdcfe9bcce00532e09b0b95.

Andrei Cristian Popescu, Haitz Sáez de Ocáriz Borde, and Pietro Liò. Adaptive depth in looped transformers: Diagnosing learned halting gates and trajectory readouts, 2026. URL https://arxiv.org/abs/2607.20519.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.378. URL https://aclanthology.org/2023.findings-emnlp.378/.

Qwen Team. Qwen3-4B-Instruct-2507. Hugging Face model card, 2025. Model card; experiments use revision cdbee75f17c01a7cc42f958dc650907174af0554.

Sapient Intelligence. HRM-Text-1B. Hugging Face model card, 2026a. Model card; experiments use revision 1f82ac2b71222f0c100a224a33f24b44a3000b6d.

Sapient Intelligence. HRM-Text: Official implementation. https://github.com/sapientinc/HRM-Text, 2026b.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. Reasoning with latent thoughts: On the power of looped transformers. In *International Conference on Learning Representations*, volume 2025, pp. 14855–14881, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2676109d49d1eb26d6bc584a8f556305-Abstract-Conference.html.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17456–17472. Curran Associates, Inc., 2022. doi: 10.52202/068431-1269. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6fac9e316a4ae75ea244ddcef1982c71-Abstract-Conference.html.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, volume 2025, pp. 10131–10165, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

TOMG Group, University of Maryland. Huginn-0125. Hugging Face model card, 2025. Model card; experiments use revision bb6621b65e90b6a4b9b29ef88dc83866d450470c.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. Hierarchical reasoning model, 2025. URL https://arxiv.org/abs/2506.21734.

Guan Wang, Changling Liu, Chenyu Wang, Cai Zhou, Yuhao Sun, Yifei Wu, Shuai Zhen, Luca Scimeca, and Yasin Abbasi Yadkori. HRM-Text: Efficient pretraining beyond scaling, 2026. URL https://arxiv.org/abs/2605.20613.

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

Hugh Mee Wong, Rick Nouwen, and Albert Gatt. When models decide and when they bind: A two-stage computation for multiple-choice question answering. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 22804–22821. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1144. URL https://aclanthology.org/2026.findings-acl.1144/.

Jian Wu, Linyi Yang, Zhen Wang, Manabu Okumura, and Yue Zhang. CofCA: A STEP-WISE counterfactual multi-hop QA benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 14631–14649, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2628d4d3b054c2d7ad33ab03435204f4-Abstract-Conference.html.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10210–10229. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.550. URL https://aclanthology.org/2024.acl-long.550/.

Sohee Yang, Nora Kassner, Elena Gribovskaya, Sebastian Riedel, and Mor Geva. Do large language models perform latent multi-hop reasoning without exploiting shortcuts? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3971–3992. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.205. URL https://aclanthology.org/2025.findings-acl.205/.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Jungmin Yun, June Hyoung Kwon, and Youngbin Kim. CRiT-QA: Evaluating multi-hop reasoning with counterfactual chains and distractor traps. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 5246–5255. ELRA Language Resource Association, 2026. doi: 10.63317/2jvxj7kwecuo. URL https://aclanthology.org/2026.lrec-1.410/.

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, 2025. URL https://arxiv.org/abs/2510.25741.

APPENDICES

A GUIDE TO EXPERIMENTS AND ANALYSES

This appendix maps each result to its evaluation split and role. We interpret swap and repair only when the designated final readout passes the unchanged-task competence check; a failed check makes the assay uninformative, not a zero effect. Within a system, K indexes increasing computation, but K values are not comparable across systems. For HRM, H1 is the retained intermediate state and H2 is the native final output. A query arm selects one of the two focal sources, and candidate order selects which answer is printed first.

How to read the appendices. Appendix B defines the readouts; Appendix C works through one graph edit; Appendices D–G define the data, scoring, and inference; and the remaining appendices give the results, controls, and limits. Table 2 maps the main evidence.

Evidence and units	Evaluation split and role
Ouro fictional panel: 288 balanced worlds	Tests acquisition across native K1–K4 exits without known facts. The main estimate uses all worlds; a sensitivity removes the 48 competence worlds.
LoopUS-8B fictional panel: 288 worlds	Tests sharpening across separate K1, K2, K4, and K8 runs. Saved scores compare raw-margin growth with common logit scaling.
Nanbeige4.2-3B: 48 competence and 240 intervention worlds	Tests native fixed-depth one- and two-pass outputs; only K2 is the checkpoint's trained default endpoint.
Phi-4 LoopUS: 48 competence and 240 intervention worlds	Evaluates four fixed-depth outputs; supported margin growth and choice decline form the descriptive divergence profile.
HRM-Text-1B linear panel: 48 competence and 240 intervention worlds	Compares retained H1 with native H2 after H2 passes competence. Both states come from one run and use the released H2-trained head.
Ouro natural paths: 124 2Wiki and 95 MuSiQue pairs	Tests whether complete-answer and first-divergence path effects grow across native exits.
LoopUS-1.7B natural paths: 93 2Wiki and 95 MuSiQue pairs	Repeats the endpoint comparison across separate fixed-depth runs.
Ouro and HRM three-edge panels: 24 development, 48 competence, and 144 intervention worlds per model	Tests whether the transition extends to a third relation using independently generated names and targets.
HRM branching panel: 24 development, 48 competence, and 144 intervention worlds	Tests whether the requested relation controls which of two outgoing links is followed.
Ouro format study: 24 competence and 48 intervention worlds	Tests acquisition across arrows, labeled tuples, compact JSON, and mapping sentences.
On-path, off-path, and locality controls: 24 development and 144 intervention worlds	Tests whether Ouro responds specifically to the answer-bearing path and whether graph connectivity survives a conflicting proximity cue.
Standard Qwen3-4B: 288 worlds and 36 layers	Applies the final normalization and head after every layer. Tests when control becomes readable across untied layers; only layer 36 is a native output.
Training-free Qwen: 288 worlds, including a 48-world competence subset	Tests separate K1, K2, K4, and K6 wrapper runs without recurrent training; K1 reproduces the unmodified model.
Weight sharing and supervision: two architectures, four schedules, eight seeds per cell	Each run uses 600 held-out worlds. Tests parameter tying and direct supervision at K1.
Separately fitted HRM and graph-task readouts	Tests whether simple output-head mismatch explains weak early control. These readouts measure linear decodability, not native behavior.
Output-interface and applicability analyses	Tests displayed-answer removal, free generation, token-boundary choices, and settings in which competence is too low to interpret the intervention.

Table 2: **Evidence map.** Where required, intervention claims follow the stated competence check. Repeated states, questions, orders, prompts, and readouts remain grouped within pair, world, or seed.

B MODEL READOUTS AND THE MEANING OF DEPTH

Depth contrasts are within systems. Equal K or H values do not imply equal computation, parameters, or training across architectures. A *native output* comes from the released model's ordinary forward and output head; a *trained exit* was explicitly supervised or designated during training. These properties differ: Nanbeige K1 is a native fixed-depth one-pass output, but only K2 is its trained default endpoint. Table 3 defines each depth index and its execution check.

System	Added computation and readouts	Precision	Execution check
Ouro-2.6B / 1.4B	Weight-tied recurrent core; native K1–K4 exits	bfloat16	Forced-exit logits match a true K-step truncation on the first prompt of each run.
LoopUS-1.7B / 8B / Phi-4	Encoder, gated reasoning block, and decoder; fixed-depth K1–K8 forwards	bfloat16	Instrumented block count equals requested K for every row.
Nanbeige4.2-3B	Shared 22-layer stack; native fixed-depth one- and two-pass outputs	bfloat16	Built-in loop count requests one or two passes; repeated K2 logits match exactly.
HRM-Text-1B	Two slow H cycles, each containing three fast L cycles; retained H1 and native H2	bfloat16	Retained-H2 logits match the stock 65,536-way output on every prompt.
Standard Qwen3-4B	36 untied decoder blocks; fixed terminal head projected at L1–L36	bfloat16	Layer-36 A/B pair logits equal the ordinary final output exactly.
Training-free Qwen wrapper	Damped tied-window refinement; separate K-dependent runs	float32	Wrapped K1 matches the unmodified model on the first prompt of each run.
Huginn-0125	Recurrent latent block; original-only competence screens at K16 and K32	bfloat16	Manual K4 trajectory matches the native full-vocabulary output exactly.
Graph-memory model	Four matched attention and feed-forward applications	float32	Tied and untied cells match operation count and tensor shapes.

Table 3: Depth is defined within each system. Qwen layers L1–L35 are fixed-head projections rather than native exits. Training-free Qwen changes its update with K, so each K is a separate analysis rather than a nested exit.

Ouro. Ouro-2.6B revision `1ed04250da1a` is a four-step recurrent model with width 2,048 and vocabulary size 49,152 (Zhu et al., 2025; ByteDance, 2025). Ouro-1.4B revision `574fa66cb8bf` appears only in an applicability analysis. With fixed weights, `exit_at_step=K-1` requests native exit K. On each run's first prompt, all 49,152 K1–K4 forced-exit logits matched a separate K-step execution within absolute tolerance 5×10^{-4} and relative tolerance 10^{-5}. This rules out stale states from a completed four-step forward. Because Ouro was trained at several exits, the timing describes this checkpoint, not the architecture alone.

LoopUS. LoopUS separates a decoder backbone into an encoder, a selectively gated reasoning block, and a decoder (Park et al., 2026a). The natural transfer uses the author-released Qwen3-1.7B checkpoint at revision `469f03291c3b`; the first fictional panel uses Qwen3-8B revision `4424cfd8e36c`; and the additional fictional test uses Phi-4 LoopUS revision `c8e514aef96a` (Park et al., 2026b;c;d). Runs use the pinned implementation at revision `dd655b77861d`. Each K is a separate fixed-depth forward with adaptive stopping disabled; runtime counters verify K reasoning-block applications. We retain K1–K8 for 1.7B and K1, K2, K4, and K8 for the 8B and Phi-4 checkpoints. We verified the implementation, checkpoint files, prompt boundary, and executed block count, but did not reconstruct the training recipes. The evidence is specific to these releases.

LoopUS K1 includes the encoder, one reasoning-block pass, and decoder; Ouro K1 is its earliest native recurrent exit. Architecture and depth training also differ, so K values are compared only within a system.

Nanbeige. Nanbeige4.2-3B revision `b82e54bd6097` applies a shared 22-layer stack twice by default (Nanbeige, 2026). Its released configuration exposes the loop count directly. We set that count to one or two and use the ordinary final normalization and language-model head, yielding native K1 and K2 outputs. K2 is the checkpoint's trained default endpoint; K1 is a direct one-pass output but was not trained as an intermediate exit. The checkpoint has no intermediate loop loss. Both passes use the same graph prompt and weights. The 48-world competence check and 240-world intervention test use the released chat template with thinking disabled and one-token terminal-code continuations. Repeated K2 forwards produce identical full-vocabulary logits.

HRM-Text. HRM-Text-1B revision `1f82ac2b71222f0c100a224a33f24b44a3000b6d` has two slow H cycles, each with three fast L cycles (Wang et al., 2025; 2026). We retain H1 and H2 from one unmodified forward pass. H2 is native; for analysis, we also apply the released H2-trained head to H1. Before comparing the two states, retained H2 must match all 65,536 stock logits within absolute tolerance 10^{-4} and relative tolerance 10^{-5}. Every linear, branching, and three-edge prompt passed this check. The implementation and checkpoint are cited separately (Sapient Intelligence, 2026b;a).

HRM uses a prefix language-model input (PrefixLM), which conditions next-token prediction on the complete prompt prefix. Its exact contract is given in Appendix F.2. Linear and branching prompts follow their respective graph grammars but share the mask, one-token A/B boundary, H-state extraction, and H2 verification. Thus H2 is native, whereas H1 is a fixed-head intermediate readout from the same trajectory. H2-minus-H1 measures the change in graph-directed control readable through the released head. It does not identify when the relevant relation was first represented internally.

The calibration analysis fits a separate standardized linear classifier with a bias term at H1 and H2. Six folds train on original and swap rows from five relation families with fixed $L_2 = 1$ regularization, then test all three states in the held-out family; repair is never used for fitting. Because the classifiers are fit and standardized separately, their margin scales are not comparable. Results and the limits of this linear readout test appear in Appendix I.4.

Standard untied Qwen. Qwen3-4B-Instruct-2507 revision `cdbee75f17c0` is an ordinary 36-block untied Transformer (Qwen Team, 2025). One forward retains each block's final-token state, then applies the released terminal normalization layer (RMSNorm) and output head. Layer-36 A/B logits equal the ordinary output exactly, so layer 36 is native; layers 1–35 are fixed-head projections, not trained outputs. They index ordinary feed-forward depth rather than recurrent inference-time computation. Separately fitted layerwise classifiers are exploratory because their onset and magnitude depend on optimization. Appendix L.3 reports both analyses.

Huginn. The natural-panel applicability check uses `tomg-group-umd/huginn-0125` revision `bb6621b65e90` (Geiping et al., 2025; TOMG Group, University of Maryland, 2025). It has a four-layer recurrent latent block between two-layer prelude and coda modules and mean recurrence 32. A manual K4 trajectory exactly reproduced the native full-vocabulary output. This output-equivalence check was performed only at K4. The K16 and K32 competence screens use separate native runs and separate sets, so they do not form a paired depth contrast. Appendix N reports why the intervention assay was not run.

Training-free wrapper. The training-free analysis uses Qwen3-4B-Instruct-2507 revision `cdbee75f17c0` with wrapper revision `2b20a873f05b` (Qwen Team, 2025; Chen et al., 2026). Layers 15–18 are reused with anchor coefficient 0.5 and update size $1/K$. Since K changes the update, this tests only whether more refinement can produce a positive increment without recurrent training. At K1, all 151,936 wrapped and unmodified logits were compared on the first prompt of the competence run and eight intervention runs. All nine matched within absolute tolerance 10^{-4} and relative tolerance 10^{-5}.

Graph-memory model. The controlled graph-memory experiment removes pretrained-model differences. A query attends to shuffled memory with four two-edge paths and terminal self-loops. Tied models apply one block four times; untied models apply four initially identical blocks once, giving both four attention and feed-forward applications. We cross both architectures with K4-only, K3–K4, K2–K4, and K1–K4 supervision; loss weights sum to one within every schedule. Models use width 64, four heads, zero dropout, and AdamW for 1,000 updates. Tied and untied models have 38,080 and 138,880 parameters, respectively, so computation but not parameter count is matched. Each

architecture-by-schedule cell has eight training seeds and 600 held-out worlds per seed; development seed 17 is excluded.

For the depth-specific analysis, the final-only tied and untied conditions are retrained because the original weights were not retained. Separate standardized linear ridge readouts with bias terms are fit at K1–K4 on data used for neither model training nor evaluation. Appendix L.2 gives the resulting readout depth profiles and their scope.

Scoring interface. Appendix F gives the shared prompt, tokenization, cache, and candidate-scoring conventions. These conventions support within-system contrasts without treating depth as a common unit across models.

C WORKED EXAMPLE: FROM GRAPH EDIT TO PATH EFFECT

World `NPC0001` illustrates how the depth effect is computed. It uses relation pair `ROUTES_TO/ENDS_AT`, two focal paths ending at red and blue, and two disconnected paths ending at green and black. Table 4 lists the records in prompt order. Bridge swap exchanges only the focal first-hop objects; repair keeps those hops and exchanges their terminal bindings, restoring the source-to-color mapping. Appendix G gives the formal estimands and inference procedure.

Row	Original	Bridge swap	Coherent repair
1	`LJOSDB ROUTES_TO HNMFCL`	`LJOSDB ROUTES_TO WOCPIX`	`LJOSDB ROUTES_TO WOCPIX`
2	`SQITSA ROUTES_TO WOCPIX`	`SQITSA ROUTES_TO HNMFCL`	`SQITSA ROUTES_TO HNMFCL`
3	`YBGIWJ ROUTES_TO UBQPNF`	same	same
4	`EDDVBH ROUTES_TO LHYFWR`	same	same
5	`UBQPNF ENDS_AT green`	same	same
6	`LHYFWR ENDS_AT black`	same	same
7	`HNMFCL ENDS_AT red`	same	`HNMFCL ENDS_AT blue`
8	`WOCPIX ENDS_AT blue`	same	`WOCPIX ENDS_AT red`

Table 4: Bridge swap changes only the focal first hops; repair restores the endpoints by changing terminal bindings. The stored prompts include the `SRC_` and `MID_` prefixes omitted here. "Same" denotes text identical to the original cell.

The user message calls the records an invented directed graph, prints the eight triples, and ends with this first-source query:

```
QUESTION
Starting from SRC_LJOSDB, follow ROUTES_TO once and then ENDS_AT once.
Which candidate terminal code is reached?

CANDIDATES
A. red
B. blue

Return exactly A or B and nothing else.
ANSWER =
```

The system message requires using only this graph and returning one label. Each world evaluates both source queries and both candidate orders. The second query replaces `SRC_LJOSDB` with `SRC_SQITSA`, and the reversed order lists blue first. Thus each focal color and each A/B label is correct equally often within the world.

Table 5 gives Ouro-2.6B's scores. For each state, $\bar{z}^{\mathrm{ref}}$ is the reference-minus-other logit margin averaged over queries and orders. "Choice" averages credit of one for a reference win, one-half for a tie, and zero for a loss. At exit K, D_K and C_K subtract swap from original; a depth gain subtracts the K1 effect from the later effect. The margin-recovery ratio R_K^m divides repair-minus-swap by

original-minus-swap margin. All retain the original answer as reference and are therefore not state accuracy.

	Original		Swap		Repair		Path effect	
K	$\bar{z}^{\text{ref}}$	choice	$\bar{z}^{\text{ref}}$	choice	$\bar{z}^{\text{ref}}$	choice	D_K	C_K
1	-0.016	0.500	-0.023	0.500	0.000	0.500	0.008	0.000
2	0.016	0.500	-0.031	0.500	0.000	0.500	0.047	0.000
3	0.172	0.500	-0.109	0.375	0.047	0.500	0.281	0.125
4	0.203	0.750	-0.125	0.250	0.078	0.500	0.328	0.500

Table 5: For this world, the path effect is near zero at K1 and larger by K4. Margins are in logit units; values are rounded to three decimals.

For this world, $D_4 = 0.203 - (-0.125) = 0.328$ logit units and $C_4 = 0.750 - 0.250 = 0.500$. The K1 values are 0.008 and zero, so $\Delta D_{4,1} = 0.320$ and $\Delta C_{4,1} = 0.500$. Repair moves the K4 margin from -0.125 to 0.078, giving $R_4^m = 0.203/0.328 = 0.619$. This one-world ratio only illustrates the calculation. Reported estimates divide the world-average repair-minus-swap difference by the world-average original-minus-swap difference. Each bootstrap draw first averages both differences over resampled worlds and then forms their ratio. Whole worlds are resampled within relation families.

D Constructing the natural path panels

We formed paired interventions from natural questions using model-independent graph, text, and tokenizer criteria. Each unit joins two answerable chains, $S_1 \xrightarrow{r_1} B_1 \xrightarrow{r_2} A$ and $S_2 \xrightarrow{r_1'} B_2 \xrightarrow{r_2} A'$, and keeps both benchmark questions and answer identities A, A'. Four standardized triples replace the original paragraphs, isolating annotated relations while retaining natural questions and entities.

Bridge swap exchanges B_1 and B_2 in the first-hop triples. Repair keeps those crossed hops and exchanges their terminal objects, restoring the original source-to-answer mapping. Questions, candidates, relation and entity inventory, record positions, and scored-token counts remain unchanged. Both questions and both candidate orders are evaluated in all three states. Candidate strings have equal, nonzero token counts within a pair. Appending either candidate leaves the prompt-prefix tokenization unchanged across graph states and candidate orders.

Property	2WikiMultiHopQA	MuSiQue
Source material	Official validation split	Official answerable train and development splits, used only for evaluation construction
Strict chain	One directed, non-temporal two-triple path	Exactly two steps, with the answer to step 1 used in step 2
Pairing stratum	Exact ordered (r_1, r_2) path	Canonical terminal relation r_2
Fixed source panel	124 pairs; 31 paths; 248 unique sources	95 pairs; 30 terminal relations; 108 full paths; 190 unique sources
LoopUS held-out test	93 pairs; all 31 paths	All 95 pairs; all 30 terminal relations
Ouro prompts and scored rows	1,488; 5,952 scored K1–K4 rows	1,140; 4,560 scored K1–K4 rows

Table 6: The natural panels preserve benchmark questions and entities while intervening on their relational paths. Graph state, question arm, candidate order, and exit are repeated measurements within a matched pair.

2WikiMultiHopQA. From all 12,576 validation examples (Ho et al., 2020), we retain a chain only if its two annotated triples follow $S \to B \to A$, the three canonical entities are distinct, and the annotated answer is A. We exclude temporal relations; questions without exactly one source mention; questions containing their bridge or answer; non-entity answers; and chains lacking unique alignment

to two supporting records. This yields 2,494 strict chains and 2,446 after holding out 48 sources used during development. We also exclude the 3,040 identifiers used in an earlier 2Wiki comparison set.

Pairs share an exact relation path, have distinct token sequences, avoid cross-question entity leakage, and never reuse a chain. Fifty-one paths admit a pair, but only 31 admit four nonoverlapping, tokenizer-valid pairs. The final panel therefore contains 124 pairs; a 32nd path did not support four eligible pairs. Within each path, orientation and evidence order fill each cell of a 2×2 counterbalance once.

We split the 124 pairs with a deterministic hash recorded in the released manifest into four 31-pair groups that each cover every path. The LoopUS development analysis used group 0; the held-out test used the other 93 pairs, three per path. The held-out test required K8 unchanged-graph accuracy of at least 0.60, a bootstrap lower bound above 0.50, and at least 0.55 accuracy for each query arm and candidate order. Appendix H reports the result.

MuSiQue. MuSiQue supplies explicit decomposition steps and constituent single-hop questions (Trivedi et al., 2022). From 22,355 answerable train and development rows, we require exactly two single-arrow steps, exact consumption of step 1 by step 2, distinct source, bridge, and answer entities, alignment to two supporting paragraphs, one source mention, and no bridge, answer, or answer alias in the question. This yields 1,414 strict chains and, after pairwise token, leakage, entity, and component checks, 17,848 candidate pairings.

Selection uses no model scores. A binary optimization forbids reuse of any decomposition step, covers every pairable terminal relation, and caps each relation at 18 pairs. It first maximizes panel size, yielding 95 pairs across all 30 eligible terminal relations, and then maximizes full-path coverage, yielding 108 paths. The selected panel contains 380 distinct steps from 96 train and 94 development sources. We solve with SciPy/HiGHS using presolve and zero relative optimality gap. A deterministic ranking recorded in the released manifest selects among panels attaining both objectives. A second score-independent hash assigns orientation and evidence order, with 24, 24, 24, and 23 pairs in the four cells.

Dependence and inference. Each source appears once, but answers and intermediate entities can recur. 2Wiki has 29 repeated normalized entities touching 37 pairs; MuSiQue has 30 touching 46. Connecting pairs that share any normalized entity gives 96 2Wiki and 56 MuSiQue components, the largest containing 12 and 30 pairs. The primary bootstrap samples paths then pairs for 2Wiki and whole pairs for MuSiQue. An entity-component bootstrap keeps both depth effects positive and both repair intervals above one half. This dependence stress test does not imply that pretraining knowledge of natural entities is absent.

E CONSTRUCTING CONTROLLED FICTIONAL WORLDS

The fictional panel removes real-world entity and relation facts and prevents source or bridge identifiers from recurring across worlds. It still uses familiar colors, relation-like strings, and instruction syntax, so conclusions concern arbitrary bindings under this grammar. Each world contains four separate paths $S_j \xrightarrow{r_1} B_j \xrightarrow{r_2} T_j$. Paths 1 and 2 are focal; paths 3 and 4 are disconnected distractors. Source and bridge identifiers have six-letter random suffixes and occur in exactly one world. Terminals are red, blue, green, and black.

Relation families. Six relation pairs are crossed with the 12 ordered pairs of distinct focal colors and four replicates, yielding $6 \times 12 \times 4 = 288$ worlds and 48 worlds per family:

`ROUTES_TO/ENDS_AT`	`LINKS_TO/RETURNS`
`POINTS_TO/YIELDS`	`CONNECTS_TO/RESOLVES_TO`
`FIRST_HOP/SECOND_HOP`	`MAPS_TO/TERMINATES_AT`

Separate rotation and reversal schedules for first- and second-edge records prevent any physical slot from identifying the focal route. The schedules use no model output.

Graph states and balance. If original focal paths end at T_1, T_2, bridge swap exchanges first-hop targets, sending S_1 to T_2 and S_2 to T_1. Repair keeps those hops but exchanges terminal objects, restoring each source's endpoint. Distractor records never change. Crossing three states, two queries, and two candidate orders gives 12 prompts per world and 3,456 fixed prompts. Each ordered focal-

color pair occurs 24 times, every color appears 72 times in each focal role, and each query arm and candidate order contributes 1,728 cells. "Original" names the reference orientation; no graph state describes a factual world.

Required property	Validator result
Unique cell identifiers and prompt keys	3,456 unique cells; exactly 12 factor combinations per world
No source or bridge reused across worlds	2,304 distinct opaque identifiers
Original and repair support the reference endpoint	Passed both queries and both orders in every world
Bridge swap reverses the supported endpoint	Passed all 1,152 swap cells
Distractors stay disconnected; inventory is preserved	Passed every state; identical entity, relation, and terminal multisets
Candidate boundary is tokenizer-valid	A/B, or Nanbeige's terminal codes, are stable one-token continuations for every evaluated checkpoint

Table 7: Every fictional world passes graph, balance, and tokenizer checks before scoring. If any check fails, construction stops rather than removing the affected cell.

Independent of the generator's closed-form label, the validator maps each (subject, relation) key to one object, follows the named relations from the query source, requires a focal-candidate endpoint, and maps its color to the displayed answer. A separate tokenizer preflight checks the assistant boundary and the candidate prefix under teacher forcing, where preceding candidate tokens are supplied rather than generated. It does not check graph correctness.

Competence tests and held-out controls. Each checkpoint's designated final readout is first tested on eight deterministically selected original worlds per relation family. Competence requires point accuracy at least 0.65, a world-bootstrap lower bound above 0.50, and accuracy at least 0.55 in every query arm and candidate order (Table 9); swap and repair stay unexamined until it passes. The main Ouro and LoopUS-8B estimates use all 288 worlds; a sensitivity analysis removes the 48 worlds used for the original-only competence check. HRM, Nanbeige, and Phi-4 LoopUS use those 48 worlds only for competence and estimate intervention effects on the remaining 240, exactly 40 per relation family. Structural development uses 24 separate execution-check worlds. Its test uses 144 others, exactly two per relation-family and ordered-color-pair combination; the other 120 of 288 worlds are unused.

Natural panels connect to benchmark entities and relations; fictional worlds better control memorized truth and repeated identity.

E.1 Branching-task construction

The branching task changes composition, not only prompt format. Relation r_1 selects one bridge from each source; each bridge has two outgoing edges, distinct relations r_X and r_Y, to plausible terminal codes. The query names the outgoing relation. Focal first-hop swap reverses the endpoint; topology repair keeps the swap but exchanges the focal bridges' terminal mappings, restoring it. Because both second relations occur at every bridge, success requires composing the selected bridge with the requested branch.

The 192 new worlds reserve 48 original graphs for H2 competence. The 144 intervention worlds form an exact $6 \times 12 \times 2$ cross of relation family, ordered focal terminal pair, and replicate. Three states, two source arms, two requested relations, and two orders give 24 prompt cells per world; H1 and H2 read the same cells. Replicates reverse which relation record is near the queried first hop, while schedules balance the queried source, requested relation, terminal identity, A/B label, record order, and locality. These worlds overlap with neither competence nor the 24 development worlds. An independent solver validates every transformation and label before scoring.

E.2 PROMPT-FORMAT STUDY

The format study asks whether Ouro's acquisition depends on this grammar. Its 72 worlds do not overlap the base panel: 24 are reserved for competence testing and 48 for effect estimation. Each world has four equivalent forms with fixed graph, query, graph-supported answer, and intervention. For each form, source and bridge names come from the same role-blind symbol pool: spelling does not reveal a node's role.

Factor	Levels	Balance
Serialization	Arrows, tuples, compact JSON, mapping sentences	All four render the same world
Opaque names	Uppercase with underscore, lowercase with underscore, title case joined, uppercase joined	Same pool for source and bridge; identifiers globally unique
Question wording	Four phrasings	Rotated independently of serialization
Evidence order	Eight rotation/reversal schedules	Each appears six times per serialization
Within-world factors	Original and bridge-swap states, query arms, and candidate orders	Fully crossed within every surface

Table 8: The format study preserves each graph and intervention across four surface forms. All 16 morphology-by-wording combinations occur three times among the 48 intervention worlds.

The competence test scores intact K4 prompts from the 24 reserved worlds. It requires overall accuracy at least 0.65, a world-bootstrap lower bound above 0.50, and accuracy at least 0.55 in every serialization, query-arm, and candidate-order slice; failure precludes swap scoring. The 48 intervention worlds yield 1,536 prompts. Analysis averages arms and orders within each surface, then the four surfaces within a world. The 10,000-draw percentile bootstrap resamples whole worlds within 16 morphology-by-wording combinations, coupling every rendering and factor cell. An exact graph solver verifies the graph-supported answer, swap reversal, preserved inventory, and identifier uniqueness.

F PROMPT CONTRACTS AND CANDIDATE SCORING

This section specifies the prompt templates and scoring interfaces. The released artifact contains byte-exact serialized prompts; line wrapping in the displays adds no characters. Each checkpoint uses its specified tokenizer and chat template, and every depth receives the same prompt.

F.1 NATURAL AND FICTIONAL BRIDGE PROMPTS

The natural assay uses one user message and no added system message. A complete 2Wiki original cell is:

```
Choose the correct answer entity by following the two linked evidence triples.

EVIDENCE
SUBJECT = Hey Gyp (Dig the Slowness) | RELATION = composer | OBJECT = Donovan
SUBJECT = Right to Dream | RELATION = composer | OBJECT = Mariah Carey
SUBJECT = Donovan | RELATION = country of citizenship | OBJECT = British
SUBJECT = Mariah Carey | RELATION = country of citizenship | OBJECT = America

QUESTION
What nationality is the composer of song Right To Dream?

CANDIDATE ANSWERS
1.  America
2.  British

Return exactly one candidate entity and nothing else.
ANSWER =
```

Only relational assignments change. Bridge swap exchanges the two first-hop objects, changing the answer from America to British. Repair keeps the crossed first hops and exchanges the second-

hop answer assignments between bridge subjects, restoring America. The question, record order, candidates, instructions, and answer boundary do not change. MuSiQue substitutes its natural question and four standardized triples. Query and order variants change only the question or candidate-line order.

Ouro's primary natural score begins with one ASCII space and the candidate as printed: `␣America`, not `America`. LoopUS instead scores the bare candidate at the Qwen assistant boundary. We tokenize each complete continuation and exclude the end-of-sequence (EOS) token. Multi-token candidates are scored under teacher forcing: each token is evaluated with the preceding candidate tokens supplied, rather than generated. These model-specific boundaries define the sequence margin and first-divergence choice; Appendix M.4 tests alternative boundaries for Ouro.

For most fictional tests, the system message adds: "Use only the supplied fictional directed graph. Follow the named relations exactly. Your entire answer must be one label: A or B." The user message calls all records invented, prints eight triples, asks for one named two-edge traversal, and ends "Return exactly A or B and nothing else. ANSWER =". Nanbeige instead asks for the displayed terminal code and scores its one-token continuation; the graph and candidate pair are unchanged. Its instruction ends, "Your entire answer must be one terminal code," and the return line requests exactly that code. Thinking is disabled. Appendix C gives an A/B example. An independent solver supplies every supported answer.

F.2 HRM BRANCHING PROMPT AND PREFIXLM READOUT

HRM-Text uses a prefix language-model input (PrefixLM): the complete instruction, examples, and target prefix condition the next-token answer. The input begins with direct-condition token `<|object_ref_start|>`, followed by the system message, two solver-verified branching examples, and the target. The examples have opposite A/B answers and use identifiers and relations absent from the test panel. Every input token has `token_type_id=1`.

The system message is "Use only the supplied fictional directed graph. Follow both named relations exactly. Your entire answer must be one label: A or B." The display below shows the target user message for one original-state test. The released prompt manifest contains the byte-exact system message, both demonstrations, the target, special tokens, and token-type identifiers.

```
This is a fictional directed graph.  Every node and record is invented.  Use only the
records below.

RECORDS
SUBJECT = SRC_2EC52E80AB | RELATION = ROUTES_TO | OBJECT = MID_A9839799C5
SUBJECT = MID_9360B51FB3 | RELATION = RETURNS | OBJECT = green
SUBJECT = SRC_5F54344FCD | RELATION = ROUTES_TO | OBJECT = MID_AE77CCCF2B
SUBJECT = SRC_C2A8198434 | RELATION = ROUTES_TO | OBJECT = MID_9360B51FB3
SUBJECT = MID_C4D217BB68 | RELATION = RETURNS | OBJECT = red
SUBJECT = MID_A9839799C5 | RELATION = RETURNS | OBJECT = blue
SUBJECT = SRC_BFA69B4AB5 | RELATION = ROUTES_TO | OBJECT = MID_C4D217BB68
SUBJECT = MID_AE77CCCF2B | RELATION = RETURNS | OBJECT = black
SUBJECT = MID_C4D217BB68 | RELATION = ENDS_AT | OBJECT = blue
SUBJECT = MID_AE77CCCF2B | RELATION = ENDS_AT | OBJECT = green
SUBJECT = MID_9360B51FB3 | RELATION = ENDS_AT | OBJECT = black
SUBJECT = MID_A9839799C5 | RELATION = ENDS_AT | OBJECT = red

QUESTION
Starting from SRC_2EC52E80AB, follow ROUTES_TO once and then ENDS_AT once.  Which
candidate terminal code is reached?

CANDIDATES
A. red
B. blue

Return exactly A or B and nothing else.
ANSWER =
```

A and B are one-token continuations. Evaluation uses the released checkpoint without adapters or fine-tuning. Appendix B defines the H1 and H2 readouts and the retained-H2 parity check.

F.3 PROMPT FORMATS USED IN THE FORMAT CHECK

The format study fixes the system message, A/B boundary, and graph while varying record and question wording. For an edge $s \xrightarrow{r} b$, the four renderers use:

Renderer	Record form
Arrows	`S --[R]--> B`
Labeled tuples	`(source: S; relation: R; target: B)`
Compact JSON	`{"source":"S","relation":"R","target":"B"}`
Mapping sentences	`Relation R maps S to B.`

Four query templates express the same composition: start at s and follow r_1 then r_2; take r_1 once then r_2 once; follow the two-edge path via r_1 then r_2; or begin at s, traverse r_1, then r_2. Each asks which displayed endpoint is reached. Appendix E.2 gives the renderer, identifier, wording, and evidence-order schedule.

F.4 STRUCTURAL PROMPT VARIANTS

Structural tests use the same fictional format. In the on/off-path test, base and transposed prompts differ only by exchanging first-hop objects in one pair of record positions. The on-path prompt queries that component and its terminals; the off-path prompt queries an unchanged component. The same edit changes only its relation to the query.

The locality test fixes the queried first hop and varies terminal binding and position. In the matched condition, its completing fact is adjacent; in the crossed condition, the adjacent fact belongs to the competing bridge and the relevant fact is farther away. Normal and transposed mappings reverse the color, so no fixed color, label, or nearby record defines the contrast.

F.5 LIKELIHOOD AND CHOICE SCORES

For rendered prefix x and natural candidate $y = (y_1, \ldots, y_m)$, the stored exit-K score is

$$\ell_K(y \mid x) = \sum_{t=1}^{m} \log p_K(y_t \mid x, y_{<t}).$$

Both natural candidates within a row use the model-specific boundary above, are scored under the same teacher-forcing rule, and exclude EOS. Their summed-log-likelihood difference is the row margin in nats, not a single-token logit difference. Candidates have equal nonzero token counts within a pair, though counts can differ across pairs.

Choice compares the candidate logits at the first position where their tokenizations differ. The preceding context is identical, so the sign is invariant to common positive rescaling. Within the secondary choice analysis, the first-divergence rule was fixed before the scores were examined; the sign of the full-sequence margin is a sensitivity. Fictional prompts compare two one-token logits: A/B for most systems and the two displayed terminal codes for Nanbeige. Positive, negative, and zero signs receive one, zero, and one-half. Appendix M.4 tests bare-token and combined boundaries. All scoring uses fixed weights with gradients and the key-value cache disabled. The target answer is not appended to the scored prompt; both candidates follow the same teacher-forcing rule.

F.6 CONTROLLED GRAPH-MEMORY INPUT

The controlled graph-memory experiment uses no text. Each example has a source query, one of four terminal labels, and twelve shuffled subject-object pairs: four edges of each type, source-to-bridge, bridge-to-terminal, and terminal self-loop. At each block, the query attends to the same memory and produces terminal logits. Tied cells reuse one block; untied cells use four copies that are identical only at initialization. We compare supervision at K4 only, K3–K4, K2–K4, and K1–K4; losses receive equal weight within each schedule. Tokenization and answer length therefore cannot explain the timing.

G ESTIMANDS, DECISION RULES, AND UNCERTAINTY

This section connects prompt-level scores to the quantities reported in the paper. The independent unit is a natural question pair or a fictional world. Repeated prompts are averaged within that unit before contrasts are formed, and uncertainty is computed by resampling whole units rather than prompt rows.

G.1 FROM CANDIDATE SCORES TO PATH CONTROL

Shared linear estimands. Let $i \in \{1, \dots, n\}$ index an independent pair or world, s a graph state, a a query arm, o a candidate order, and t a reported readout. Let p_{isao} be the prompt prefix and $\ell_t(y \mid p)$ the candidate score at readout t. We orient every score toward the original graph's answer for query arm a:

$$x_{isaot} = \ell_t(y_{ia}^{\text{ref}} \mid p_{isao}) - \ell_t(y_{ia}^{\text{other}} \mid p_{isao}).$$

Natural rows use the complete-answer log-likelihood from Appendix F.5. Their secondary choice analysis uses the logit margin $\widetilde{x}_{isaot}$ at the first token where the two candidate strings differ. Fictional rows use the one-token candidate-logit difference, so $\widetilde{x}_{isaot} = x_{isaot}$. Row-level choice credit is

$$q_{isaot} = \mathbb{1}(\widetilde{x}_{isaot} > 0) + \tfrac{1}{2}\mathbb{1}(\widetilde{x}_{isaot} = 0),$$

where $\mathbb{1}(\cdot)$ equals one when its condition is true and zero otherwise. The four query-arm and candidate-order rows are averaged within a unit:

$$z_{ist} = \tfrac{1}{4}\sum_{a,o} x_{isaot}, \qquad c_{ist} = \tfrac{1}{4}\sum_{a,o} q_{isaot}.$$

Assigning credit before averaging preserves the four paired choices instead of taking the sign of their mean margin.

Write O, S, and R for the original, swap, and repair states. Because the correct answer reverses under the swap, graph-correct choice credit is

$$q_{isaot}^{\text{graph}} = \begin{cases} q_{isaot}, & s \in \{O, R\}, \\ 1 - q_{isaot}, & s = S, \end{cases} \qquad \text{acc}_{ist} = \tfrac{1}{4}\sum_{a,o} q_{isaot}^{\text{graph}}, \quad A_{st} = n^{-1}\sum_i \text{acc}_{ist}\,.$$

Thus A_{st} asks whether the readout selects the answer supported by state s, whereas the path-control contrasts below keep the original answer identity fixed across states.

Only the format study adds a rendering index $u \in \{1, 2, 3, 4\}$. In that study, we first average arms and orders within each rendering, then average renderings:

$$z_{ist} = \tfrac{1}{4}\sum_u \left(\tfrac{1}{4}\sum_{a,o} x_{isuaot}\right), \qquad c_{ist} = \tfrac{1}{4}\sum_u \left(\tfrac{1}{4}\sum_{a,o} q_{isuaot}\right).$$

The four renderings improve precision within a world; they do not increase the independent sample size.

We write D_t for the original-minus-swap score change and C_t for the corresponding change in answer choice. The Δ terms ask whether each response grows from an earlier readout t_0 to a later readout t_1. Repair increments N_t and Q_t measure the reversal from swap to repair; R_t^m and R_t^c give the recovery ratio. Their exact definitions are

$$d_{it} = z_{iOt} - z_{iSt}, \qquad D_t = n^{-1}\sum_i d_{it}, \tag{2}$$

$$e_{it} = c_{iOt} - c_{iSt}, \qquad C_t = n^{-1}\sum_i e_{it}, \tag{3}$$

$$\Delta D_{t_1,t_0} = n^{-1}\sum_i (d_{it_1} - d_{it_0}), \qquad \Delta C_{t_1,t_0} = n^{-1}\sum_i (e_{it_1} - e_{it_0}). \tag{4}$$

Repair is measured relative to the swap:

$$N_t = n^{-1}\sum_i (z_{iRt} - z_{iSt}), \quad Q_t = n^{-1}\sum_i (c_{iRt} - c_{iSt}),$$
$$R_t^m = N_t / D_t, \qquad R_t^c = Q_t / C_t.$$

Each bootstrap draw recomputes these pooled ratios. One denotes full average recovery and values above one denote overshoot. A ratio is left undefined when its stated denominator condition fails.

Natural sequence margins and fictional label logits have different units, so C_t is the common decision scale.

Branching interactions. For branching, $r \in \{X, Y\}$ denotes the requested outgoing relation. For both requests, margins are oriented toward the endpoint of the original X query, so the Y response enters the interaction with the opposite sign. Ordinary choice credit $q^{(0)}_{israot} \in \{0, \frac{1}{2}, 1\}$ is mapped to

$$q^{\pm}_{israot} = 2q^{(0)}_{israot} - 1, \qquad \bar{x}_{isrt} = \tfrac{1}{4}\sum_{a,o} x_{israot}, \quad \bar{q}_{isrt} = \tfrac{1}{4}\sum_{a,o} q^{\pm}_{israot}.$$

For either $v = \bar{x}$ or $v = \bar{q}$, the topology-repair interaction is

$$I^{\mathrm{rep}}_{it}(v) = \tfrac{1}{4}\{(v_{iRXt} - v_{iSXt}) - (v_{iRYt} - v_{iSYt})\}, \qquad I^{\mathrm{rep}}_{t}(v) = n^{-1}\sum_i I^{\mathrm{rep}}_{it}(v).$$

Replacing R by O defines the bridge interaction I^{bridge}_t. Under perfect composition, the four signed choice differences sum to four, so $I^{\mathrm{rep}}_t(\bar{q}) = 1$. At H2, choice recovery is $I^{\mathrm{rep}}_{H2}(\bar{q})/I^{\mathrm{bridge}}_{H2}(\bar{q})$. State accuracy continues to credit the endpoint supported by each graph. We test H2 composition before the H2-minus-H1 change.

Structural and training contrasts. For the on/off-path test, E_{rt} is the base-minus-transposed response when the edited component is on or off the queried path. Specificity is $S_t = E_{\mathrm{on},t} - E_{\mathrm{off},t}$, and $S_4 - S_1$ is its depth change. For locality, L_{bt} is the transposed-minus-normal terminal-binding response for $b \in \{\mathrm{matched}, \mathrm{crossed}\}$. The planned endpoint is $L_{\mathrm{crossed},4}$; matched minus crossed measures the additional association with physical locality. Both are evaluated in margin and exact choice.

In the graph-memory experiment, $g_{\tau\nu}$ is K4-minus-K1 choice gain for $\tau \in \{\mathrm{tied}, \mathrm{untied}\}$ and $\nu \in \{\mathrm{final}, \mathrm{every}\}$. We report tied minus untied at each supervision level, every-exit minus final-only within each architecture, and their difference-in-differences. Seed and evaluation world are shared across all four conditions.

G.2 Competence checks and decision rules

Unchanged-task accuracy establishes that an interface can express the task before its intervention is interpreted. A required failure stops that study; it is not reported as an intervention null. The original Ouro natural panels are the exception: their final accuracy is descriptive and was not a required check.

Natural panels. Ouro on 2Wiki and MuSiQue requires positive lower bounds for D_3 and $\Delta D_{3,1}$, positive D_3 in every bootstrap draw, and an R^m_3 lower bound above 0.50. LoopUS on 2Wiki applies the same K8-versus-K1 rules, with an R^m_8 lower bound above 0.50. LoopUS on MuSiQue requires positive lower bounds for D_8, $\Delta D_{8,1}$, and the repair increment; $R^m_8 \geq 0.50$ with a positive lower bound; and positive D_8 and depth gain in every query and order slice. Every condition must hold.

Fictional linear and format studies. Ouro acquisition on the base panel requires positive lower bounds for D_4, $\Delta D_{4,1}$, and $\Delta C_{4,1}$. The format study applies the same three conditions on its 48 intervention worlds. Nanbeige and Phi-4 use the analogous requirements at K2 versus K1 and K8 versus K1. LoopUS-8B and Phi-4 fail the frozen positive-choice-gain condition; for Phi-4, the estimated choice change is negative. Margin-control divergence is a descriptive, post-result label for Phi-4's supported margin increase and choice decrease, not a frozen decision category. HRM linear transfer requires a positive H2 path effect, positive H2-minus-H1 margin and choice changes, and a positive H2 repair increment.

Three-edge paths. Let t_0 and t_1 denote the early and late readouts. A qualifying transition requires original- and swap-state accuracy at t_1 of at least 0.60, with both lower bounds above 0.50; $C_{t_1} \geq 0.10$; $\Delta C_{t_1,t_0} \geq 0.05$; and positive lower bounds for C_{t_1}, $\Delta C_{t_1,t_0}$, D_{t_1}, and $\Delta D_{t_1,t_0}$. Repair additionally requires repair accuracy of at least 0.60 with a lower bound above 0.50, and positive lower bounds for N_{t_1} and Q_{t_1}. Choice recovery is interpreted only when C_{t_1} has a positive lower bound and at least 99% of bootstrap ratios are finite; its estimate and lower bound must both be at least 0.50.

Study	Required unchanged-task check
Ouro natural K1–K3	No required check; K3 graph-consistent accuracy and its interval are reported descriptively in Appendix H.
LoopUS natural	At K8: overall ≥ 0.60, lower bound > 0.50, and every arm/order ≥ 0.55. On 95 MuSiQue pairs, the prior competence test passed at $0.787\,[0.745, 0.829]$; its lowest arm/order accuracy was 0.758.
Fictional two-edge	On 48 original-only worlds: overall ≥ 0.65, lower bound > 0.50, and every arm/order ≥ 0.55, evaluated at Ouro K4, LoopUS K8, HRM H2, Nanbeige K2, or Phi-4 K8.
Training-free Qwen	The same fictional threshold on 48 original-only worlds at K1, after checking that the wrapper reproduces the unmodified model.
Three-edge Ouro and HRM	The same overall and arm/order thresholds on 48 original-only worlds separate from development and intervention worlds.
HRM branching	H2 overall ≥ 0.65 with lower bound > 0.50; every relation, source, and order ≥ 0.55; every relation family ≥ 0.50.
Ouro format study	On 24 original-only worlds: overall ≥ 0.65, lower bound > 0.50, and every renderer/arm/order ≥ 0.55.
Ouro structural controls	The fictional competence check above, followed by execution checks on a separate 24-world set.
Graph-memory experiment	K4 accuracy ≥ 0.95 in all 64 architecture-by-schedule runs.

Table 9: **Required competence checks before intervention testing.** Brackets are 95% percentile-bootstrap intervals.

Branching. After the competence check, H2 composition is tested first. The choice interaction must be at least 0.10 with a positive lower bound. Its raw counterpart and the X, Y, and crossed-locality repair-choice effects must also have positive lower bounds. Original accuracy must be at least 0.65, swap and repair accuracy at least 0.60, and all three lower bounds above 0.50. The choice-repair ratio and its lower bound must both be at least 0.50. Only then is depth tested: H2-minus-H1 choice gain must be at least 0.05 with a positive lower bound, and raw gain must have a positive lower bound.

Structural controls and graph-memory models. Each structural branch requires raw and choice lower bounds above zero for its planned K4 effect and K4-minus-K1 change. In the graph-memory experiment, delayed acquisition under final-only loss requires positive lower bounds on K4-minus-K1 choice gain for both tied and untied models. A tying effect requires a positive lower bound on their final-only gain difference. Earlier control under every-exit supervision requires negative upper bounds on the every-exit-minus-final-only gain in both architectures. Their difference-in-differences is secondary.

The main Ouro and LoopUS-8B estimates include all 288 worlds, including the 48 used for original-only competence screening. A sensitivity analysis excludes those 48 and uses the remaining 240. HRM, Nanbeige, and Phi-4 use only those 240 intervention worlds. The format study separates competence and intervention worlds; HRM branching uses 144 new worlds; and each three-edge study uses 48 competence and 144 different intervention worlds.

G.3 Evaluation and inference

Algorithm 1 shows the common swap-and-repair workflow. Format and structural studies change the graph states but retain the competence test and paired unit-level inference.

Bootstrap construction. Every analysis keeps all states, prompts, and readouts from a selected unit together. Table 10 gives the independent unit and any stratification used by each study.

Within each analysis, all reported 95% intervals use the same 10,000 nonparametric percentile draws. Each primary claim requires every stated one-sided condition to pass. Intervals are not adjusted across separate studies, and we make no paper-wide familywise-error claim. On the natural panels, complete-answer margin and its depth gain are primary. Within the secondary choice analysis, the first-divergence sign is the fixed choice measure; the sign of the full-sequence margin is a sensitivity. Secondary readouts, stratum summaries, residual associations, calibration checks, and other sensitivity analyses are descriptive. Root-mean-square-normalized margins provide an additional natural-panel

Algorithm 1 Evidence-control evaluation

Require: Evaluation units $\mathcal{I}$, competence units $\mathcal{G}$, graph states $\mathcal{S}$, readouts $\mathcal{T}$, and the predefined rules above
Construct every graph state and verify its supported answer with the symbolic solver
Evaluate the final-readout competence check on original prompts from $\mathcal{G}$
if any required competence condition fails **then**
    Stop without interpreting the intervention
end if
for all $i \in \mathcal{I}$, $s \in \mathcal{S}$, and $t \in \mathcal{T}$ **do**
    Score both candidates for every query arm and candidate order
    Orient margins toward the original answer; assign choice credit 1, 0, or $1/2$
    Credit the answer supported by state s for accuracy, then average prompts within i
end for
Compute path effects, later-minus-earlier changes, repair, and state accuracy
for 10,000 bootstrap draws **do**
    Resample whole units within the stated strata, keeping all paired fields together
    Recompute every contrast, ratio, and accuracy
end for
Form percentile intervals and apply the study-specific rules above

Analysis	Independent unit	Resampling scheme
2Wiki natural	Relation path, then pair	Resample paths, then whole pairs within each selected path.
MuSiQue natural	Whole pair	Resample 95 pairs; terminal-relation and entity-component schemes are sensitivities.
Fictional bridge and structure	World within family	Resample the observed world count separately in each of six relation families.
HRM, Nanbeige, and Phi-4 two-edge	World within family	Resample 40 of 240 worlds in each family.
HRM branching	World within family	Resample 24 of 144 worlds in each family, with states, relations, arms, orders, and H states coupled.
Three-edge studies	World within family	For each model, resample 24 of 144 worlds in each family with all paired fields coupled.
HRM linear classifiers	Held-out world within family	Resample 240 cross-fitted intervention worlds, conditioning on the six fitted H1 and H2 classifiers.
Standard Qwen layers	World within family	Resample 48 of 288 worlds in each family with layers and states coupled; fitted-classifier intervals condition on the folds.
Training-free Qwen	World within family	Resample 48 of 288 worlds in each family with depths, graph states, arms, and orders coupled.
Format study	World within design block	Resample three worlds in each of 16 morphology-by-wording blocks, keeping all surfaces and cells coupled.
Natural structural controls	Relation path; then pair for locality	Resample paths for on/off-path effects; for locality, resample 31 paths and four whole pairs per path.
Graph-memory experiment	Seed, then world	Resample eight seeds, then 600 worlds within seed while retaining both architectures and all four supervision schedules.
Graph-memory readouts	Seed, then world	Resample eight seeds, then 600 held-out worlds; depth-specific classifiers remain fixed.

Table 10: **Unit-level resampling preserves every paired comparison.** Readout analyses resample test units without refitting the classifier.

scale check. Appendix A records the status of each result. The 24 branching-development worlds are not pooled with competence or intervention data; branching inference is conditional on the six fixed relation families and four terminal codes.

H NATURAL PATHS: COMPLETE DEPTH PROFILES AND REPAIR

The complete depth profiles distinguish changes in the selected answer from changes in its margin. Repair and dependence checks then ask how closely each response tracks the supplied path. The Ouro panels had no required competence check. At K3, graph-consistent two-candidate accuracy pooled across the three graph states was $0.690\,[0.651, 0.728]$ on 2Wiki and $0.675\,[0.646, 0.705]$ on MuSiQue. LoopUS passed its 2Wiki K8 competence check at $0.798\,[0.739, 0.858]$; the lowest query-arm and candidate-order accuracies were 0.780 and 0.769. An earlier original-state test on all 95 MuSiQue pairs also passed: K8 accuracy was $0.787\,[0.745, 0.829]$, and the lowest query-arm or candidate-order accuracy was 0.758. Appendix G.2 gives the common thresholds.

H.1 COMPLETE DEPTH PROFILES

Table 11 gives D_K and C_K at every reported readout. Complete-answer D_K and its endpoint gain are primary; first-divergence C_K is secondary. The endpoint comparisons are Ouro K3 versus K1 and LoopUS K8 versus K1. Intermediate readouts are descriptive; LoopUS K2–K7 were added after K1 and K8 reproduced the released outputs.

2WikiMultiHopQA			
System	Readout	D_K	C_K
Ouro-2.6B	K1	$-0.001\,[-0.019, 0.014]$	$0.004\,[-0.011, 0.019]$
	K2	$1.761\,[1.494, 2.048]$	$0.190\,[0.139, 0.242]$
	K3	$3.114\,[2.734, 3.510]$	$0.388\,[0.306, 0.471]$
	K4	$3.580\,[3.138, 4.027]$	$0.432\,[0.339, 0.528]$
LoopUS-1.7B	K1	$2.066\,[1.629, 2.529]$	$0.220\,[0.138, 0.304]$
	K2	$3.164\,[2.643, 3.694]$	$0.364\,[0.293, 0.437]$
	K3	$3.185\,[2.663, 3.714]$	$0.356\,[0.284, 0.430]$
	K4	$3.187\,[2.665, 3.715]$	$0.352\,[0.281, 0.425]$
	K5	$3.182\,[2.658, 3.715]$	$0.356\,[0.284, 0.430]$
	K6	$3.184\,[2.665, 3.714]$	$0.362\,[0.290, 0.434]$
	K7	$3.177\,[2.657, 3.707]$	$0.363\,[0.290, 0.437]$
	K8	$3.174\,[2.664, 3.693]$	$0.362\,[0.289, 0.435]$
MuSiQue			
System	Readout	D_K	C_K
Ouro-2.6B	K1	$0.024\,[-0.003, 0.053]$	$0.013\,[0.000, 0.026]$
	K2	$1.864\,[1.628, 2.116]$	$0.199\,[0.157, 0.243]$
	K3	$3.165\,[2.813, 3.521]$	$0.367\,[0.303, 0.433]$
	K4	$3.497\,[3.121, 3.881]$	$0.442\,[0.367, 0.517]$
LoopUS-1.7B	K1	$1.897\,[1.578, 2.246]$	$0.213\,[0.159, 0.267]$
	K2	$2.590\,[2.236, 2.980]$	$0.291\,[0.229, 0.353]$
	K3	$2.598\,[2.242, 2.988]$	$0.286\,[0.224, 0.347]$
	K4	$2.595\,[2.239, 2.986]$	$0.292\,[0.232, 0.354]$
	K5	$2.594\,[2.241, 2.983]$	$0.287\,[0.225, 0.350]$
	K6	$2.594\,[2.239, 2.984]$	$0.296\,[0.236, 0.358]$
	K7	$2.584\,[2.231, 2.974]$	$0.295\,[0.233, 0.357]$
	K8	$2.583\,[2.224, 2.978]$	$0.299\,[0.237, 0.361]$

Table 11: **Complete natural-panel depth profiles.** D_K is the complete-answer path effect in nats, and C_K is the first-divergence choice effect. Brackets are 95% intervals, and choice profiles are secondary. Ouro rows are native exits; each LoopUS row is a separate fixed-depth output. Readout labels are comparable only within a system.

Ouro's control grows across the natural paths. At K3, original and swap margins are 2.232 and -0.882 nats on 2Wiki, and 2.518 and -0.648 on MuSiQue. Thus the separation reflects support for the original answer under the original graph and for the reversed answer after the swap.

LoopUS changes mostly between K1 and K2. LoopUS separates the states at K1. Its K1-to-K2 increments are $1.098\,[0.844, 1.365]$ nats on 2Wiki and $0.693\,[0.494, 0.907]$ on MuSiQue; every adjacent change after K3 is at most 0.011 nats. From K1 to K2, its first-divergence choice effect rises from 0.220 to 0.364 on 2Wiki and 0.213 to 0.291 on MuSiQue, then plateaus. These checkpoint-specific results do not isolate an architectural effect.

H.2 Repair and decision-level robustness

Repair recovers most of the natural effect (Table 12); Ouro K1 ratios are undefined because D_1 is not bounded away from zero. First-divergence and complete-answer choice gains are positive in every cell.

System	Dataset	Exit	R_K^m	Full-sequence-sign gain
Ouro	2Wiki	K3	$0.711\,[0.623, 0.784]$	$0.379\,[0.294, 0.466]$
	MuSiQue	K3	$0.648\,[0.561, 0.726]$	$0.363\,[0.297, 0.432]$
LoopUS	2Wiki	K8	$0.790\,[0.710, 0.859]$	$0.156\,[0.075, 0.239]$
	MuSiQue	K8	$0.670\,[0.573, 0.753]$	$0.071\,[0.003, 0.145]$

Table 12: **Repair and discrete choice control.** R_K^m is the recovery ratio after coherent repair. Full-sequence-sign gain is the late-minus-early change in the original-minus-swap choice effect computed from the sign of the complete-answer margin. Brackets are 95% intervals.

Aggregate depth-gain intervals remain positive under the alternative dependence checks. For Ouro, all 31 2Wiki relation-path means and all 30 MuSiQue terminal-relation means are positive. Resampling shared-entity components gives K3 depth-gain intervals of $[2.788, 3.443]$ nats on 2Wiki and $[2.790, 3.492]$ on MuSiQue. Removing repeated identities one at a time changes the deep effect by at most 0.048 and 0.091 nats, respectively. For LoopUS, 30 of 31 2Wiki and 20 of 30 MuSiQue group means have positive endpoint gain. Under terminal-relation resampling, LoopUS on MuSiQue retains positive path-effect and depth-gain intervals, but its recovery interval $[0.496, 0.770]$ does not clear one-half.

At the first differing token, Ouro's K3-minus-K1 gain is $0.384\,[0.301, 0.470]$ on 2Wiki and $0.354\,[0.288, 0.420]$ on MuSiQue; LoopUS's K8-minus-K1 gain is $0.141\,[0.058, 0.228]$ and $0.086\,[0.014, 0.161]$. Because each pair has the same prefix, positive temperature rescaling cannot change the sign. Root-mean-square-normalized margins and complete-answer choices agree.

H.3 Why natural repair is incomplete

The unrecovered gap, $U_K = n^{-1}\sum_i (z_{iOK} - z_{iRK})$, emerges over the same exits as Ouro's path response. At K3 it is $0.899\,[0.688, 1.130]$ nats on 2Wiki and $1.114\,[0.878, 1.355]$ on MuSiQue. To characterize this gap, we scored each terminal fact alone. Pairwise compatibility is the mean K3 log-likelihood preference, across terminal subjects and candidate orders, for a subject's original answer over the answer from its paired chain. It covaries with the residual gap: Spearman ρ is $0.458\,[0.234, 0.639]$ across 124 2Wiki pairs and $0.548\,[0.383, 0.680]$ across 95 MuSiQue pairs.

In a secondary descriptive intervention, the prompt called all supplied statements authoritative fictional facts. It increased the K3 path effect by 0.471 nats on 2Wiki and 0.309 on MuSiQue while leaving the unrecovered gap essentially unchanged: standard-minus-instructed differences were $-0.001\,[-0.061, 0.058]$ and $-0.005\,[-0.058, 0.050]$ nats. The residual covaries with learned compatibility, but this analysis does not identify mediation.

I Fictional worlds: acquisition, sharpening, divergence, and transfer

The fictional panel uses unique source and bridge identifiers (Appendix E). On 48 original-only worlds, Ouro-2.6B at K4 and LoopUS-8B at K8 passed the competence screen with accuracies $0.737\,[0.688, 0.784]$ and $0.943\,[0.909, 0.971]$. Both met the query-arm, candidate-order, and token-

boundary requirements before swap and repair were scored. Their main estimates include all 288 worlds; the 240-world exclusion sensitivity appears below.

I.1 Ouro native exits and LoopUS fixed-depth outputs

Ouro acquires graph-directed choices. Ouro starts at chance and improves at every native exit (Table 13). From K1 to K4, raw gain is $0.277\,[0.264, 0.290]$ logits and choice gain is $0.480\,[0.451, 0.509]$. Original-state accuracy rises from 0.500 to 0.734, and swap-state accuracy from 0.500 to 0.747. The K3-to-K4 choice increment is $0.212\,[0.188, 0.236]$, so the final pass still changes answers.

At Ouro K4, the repair margin is 0.142 logits, compared with 0.134 for the original and -0.142 for the swap. The repair ratio is $1.032\,[1.005, 1.058]$, and graph-correct repair accuracy is $0.740\,[0.721, 0.758]$. Repair therefore restores essentially all of the fictional swap effect; the lower natural-panel recovery is not imposed by the recovery ratio.

LoopUS sharpens a choice established early.

LoopUS-8B already follows the graph strongly at K1. From K1 to K8, the raw margin widens by $0.217\,[0.208, 0.225]$ logits while choice changes by $0.000\,[-0.010, 0.010]$. At each depth, the root-mean-square (RMS) check divides the mean path effect by the RMS of the four original-state margins in each world, two query arms by two candidate orders. The normalized K8-minus-K1 change is $-0.006\,[-0.010, -0.003]$. A second check fits the nonnegative zero-intercept map $\text{K1} = s\,\text{K8}$ on original-state margins from five relation families, then evaluates $sD_8 - D_1$ on the held-out family. Its residual is $-0.003\,[-0.009, 0.004]$. Both terms are refit in every bootstrap draw. The raw margin grows, but these checks support common logit scaling rather than an increase in choice control. At K8, original and swap accuracies are $0.931\,[0.916, 0.946]$ and $0.951\,[0.938, 0.962]$; repair accuracy is $0.934\,[0.921, 0.947]$, with $R_8^m = 1.082\,[1.061, 1.104]$. These retrospective checks address positive multiplicative scaling, not arbitrary shifts or nonlinear recalibration.

System	Readout	D_K	C_K
Ouro-2.6B	K1	$-0.001\,[-0.003, 0.001]$	$0.000\,[0.000, 0.000]$
	K2	$0.039\,[0.036, 0.043]$	$0.020\,[0.013, 0.028]$
	K3	$0.199\,[0.189, 0.209]$	$0.268\,[0.241, 0.297]$
	K4	$0.276\,[0.263, 0.289]$	$0.480\,[0.451, 0.509]$
LoopUS-8B	K1	$4.250\,[4.154, 4.345]$	$0.882\,[0.863, 0.900]$
	K2	$4.338\,[4.241, 4.436]$	$0.883\,[0.864, 0.902]$
	K4	$4.469\,[4.369, 4.570]$	$0.881\,[0.861, 0.901]$
	K8	$4.466\,[4.366, 4.567]$	$0.882\,[0.862, 0.901]$

Table 13: **Ouro acquires path control, whereas LoopUS sharpens early control.** D_K is the path effect in logits and C_K its exact-choice analogue; brackets are 95% intervals. Each LoopUS row is a separate fixed-depth run, and readout labels are comparable only within a system.

I.2 Retrospective checks of acquisition timing

We used the saved outputs from the same 288 Ouro worlds to examine whether the aggregate gain reflects stable changes within examples. This is a retrospective analysis, not a new confirmation panel. A world is positive at exit K when its original-minus-swap choice contrast, averaged over both queries and candidate orders, is greater than zero. Only 4.5% $[2.1, 6.9]$ of the resulting binary trajectories contain a later positive-to-nonpositive reversal. In 66.0% $[60.4, 71.2]$ of worlds, the contrast first becomes and remains positive at K3 (Figure 5).

We also ranked worlds by their K1 original-state absolute margin separately within each relation family, giving 72 worlds per quartile. From the least to the most confident quartile, the K4-to-K1 choice gains are $0.505\,[0.450, 0.561]$, $0.491\,[0.438, 0.543]$, $0.519\,[0.462, 0.580]$, and $0.406\,[0.345, 0.465]$. The gain is therefore not confined to worlds close to the K1 choice boundary. This observational check does not make the K1 margin a causal measure of task difficulty.

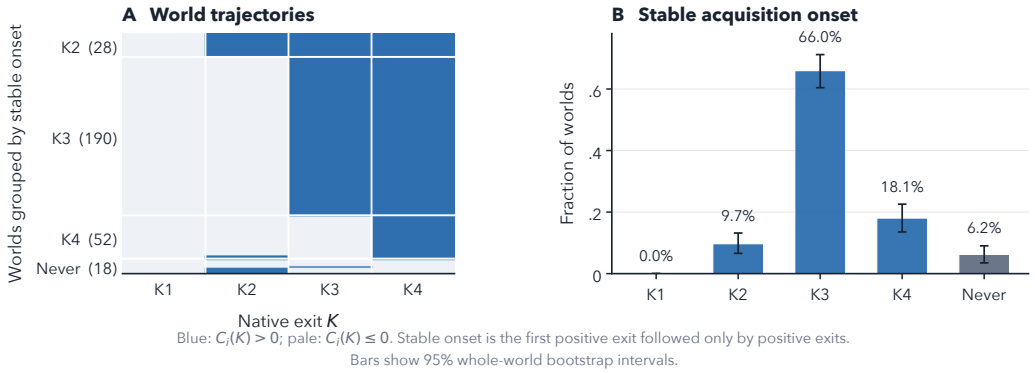

Figure 5: **Ouro's world-level choice response usually appears without later reversal. Left:** Worlds grouped by the first exit after which the original-minus-swap choice contrast remains positive. Blue cells are positive; pale cells are nonpositive. **Right:** Distribution of this stable onset. Whiskers show 95% whole-world bootstrap intervals. The analysis is retrospective and uses the frozen outputs from the main fictional panel.

I.3 Additional recurrent checkpoints

Two further checkpoints were evaluated on separate sets of 48 competence and 240 intervention worlds. Nanbeige's built-in loop-count control gives native fixed-depth one- and two-pass outputs, although only K2 is the checkpoint's trained default endpoint. Its choice effect rises from 0.020 at K1 to 0.597 at K2, a gain of $0.578\,[0.544, 0.611]$. Phi-4 LoopUS behaves differently across separate fixed-depth runs: its choice effect falls from 0.716 at K1 to 0.653 at K8, while its raw-margin effect increases by $0.035\,[0.029, 0.041]$ logits. This supported opposite-sign pattern is the descriptive margin-control-divergence profile used in the main text.

System	Readout	D_K	C_K
Nanbeige4.2-3B	K1	$0.021\,[0.018, 0.024]$	$0.020\,[0.006, 0.034]$
	K2	$6.062\,[5.801, 6.338]$	$0.597\,[0.566, 0.629]$
Phi-4 LoopUS	K1	$1.927\,[1.871, 1.984]$	$0.716\,[0.689, 0.742]$
	K2	$1.963\,[1.904, 2.021]$	$0.667\,[0.639, 0.696]$
	K4	$1.971\,[1.913, 2.029]$	$0.659\,[0.630, 0.689]$
	K8	$1.963\,[1.905, 2.021]$	$0.653\,[0.624, 0.683]$

Table 14: **Nanbeige acquires control, whereas Phi-4 exhibits margin-control divergence.** D_K is the path effect in logits and C_K its exact-choice analogue. Brackets are 95% intervals. Each Phi-4 row is a separate fixed-depth run.

Nanbeige passed its K2 competence check with accuracy $0.760\,[0.695, 0.823]$. At K2, repair restores the removed margin with ratio $1.114\,[1.066, 1.164]$ and graph-correct accuracy $0.841\,[0.819, 0.862]$. Phi-4 passed at K8 with accuracy $0.818\,[0.786, 0.849]$; its K8 repair ratio is $1.023\,[1.005, 1.042]$, with accuracy $0.829\,[0.812, 0.847]$.

I.4 HRM retained-state results

HRM-Text-1B's default two-cycle forward pass yields a retained intermediate state, H1, and its native final output, H2. We apply the released H2-trained output head to both states. The comparison therefore shows when path control becomes readable through that fixed head, not when the relevant relation is first represented internally.

Estimand	Estimate
Panel A: linear two-edge task (48 competence; 240 intervention worlds)	
H2 competence accuracy	$0.909\,[0.867, 0.948]$
H1 path-margin contrast	$0.000\,[-0.001, 0.001]$
H2 path-margin contrast	$3.768\,[3.622, 3.916]$
H2–H1 raw-margin gain	$3.768\,[3.622, 3.915]$
H2–H1 exact-choice gain	$0.896\,[0.874, 0.916]$
H2 repair increment over swap	$3.936\,[3.791, 4.084]$
H2 margin recovery R^m_{H2}	$1.045\,[1.027, 1.062]$
H2 original-state accuracy	$0.937\,[0.922, 0.951]$
H2 swap-state accuracy	$0.959\,[0.946, 0.970]$
H2 repair-state accuracy	$0.958\,[0.945, 0.969]$
Panel B: branching task (48 competence; 144 intervention worlds)	
H2 competence accuracy	$0.777\,[0.740, 0.815]$
H1 topology-repair choice interaction	$0.000\,[0.000, 0.000]$
H2 topology-repair choice interaction	$0.562\,[0.526, 0.596]$
H2–H1 choice-interaction gain	$0.562\,[0.526, 0.596]$
H2 topology-repair raw interaction	$0.867\,[0.817, 0.917]$
H2–H1 raw-interaction gain	$0.867\,[0.816, 0.917]$
H2 repair choice, requested X	$0.535\,[0.477, 0.592]$
H2 repair choice, requested Y	$0.589\,[0.531, 0.647]$
H2 crossed-locality repair choice	$0.648\,[0.589, 0.705]$
H2 choice recovery R^c_{H2}	$1.009\,[0.966, 1.052]$
H2 original-state accuracy	$0.782\,[0.763, 0.801]$
H2 swap-state accuracy	$0.775\,[0.753, 0.796]$
H2 repair-state accuracy	$0.787\,[0.767, 0.807]$

Table 15: **Path control is readable at native H2 but not through the same head at retained H1.** The branching choice interaction equals one under perfect composition. H1 is not a trained exit; brackets are 95% intervals from stratified whole-world draws with paired states.

Linear transition through the released head. HRM passed its 48-world H2 competence check before evaluation on the separate 240-world intervention set. Panel A of Table 15 shows a near-zero H1 response through the released head, strong native-H2 control, and repair. This is a retained-state-to-final-output transition, not a comparison between two trained exits.

The transition also appears on the branching task. HRM passed the 48-world H2 competence check before evaluation on the separate 144-world branching set. Panel B of Table 15 shows repair for both requested relations and under adverse locality. Every H2 composition criterion passed before the H1-to-H2 test. The result is conditional on six relation families, four terminal codes, and fixed demonstrations. Every family and leave-one-family-out estimate is positive, but these are descriptive slices rather than independent replications. The test establishes control of the predefined output comparison, not an internal graph algorithm or open-vocabulary competence.

Separately fitted linear classifiers. To test a simple head-alignment explanation, we fit a logistic classifier with a bias term at each state, using an L_2 penalty coefficient of 1. Each of six folds trains on 1,600 original and swap rows from five relation families and tests 480 original, swap, and repair rows from the held-out family. Repair is never used for fitting. Some H1 fits reached the initial 200-step limit, so we raised the limit to 1,000 and added the whole-world split reported below.

Fitted readout	C	Original acc.	Swap acc.	Repair acc.
H1 fitted classifier	$-0.004\,[-0.024, 0.015]$	$0.499\,[0.484, 0.514]$	$0.497\,[0.483, 0.510]$	$0.500\,[0.486, 0.514]$
H2 fitted classifier	$0.989\,[0.979, 0.996]$	$0.993\,[0.986, 0.998]$	$0.996\,[0.991, 1.000]$	$0.997\,[0.993, 1.000]$

Table 16: **Separate linear classifiers preserve the H1-to-H2 difference.** Brackets are 95% intervals from 10,000 stratified whole-world draws without refitting. This analysis measures linear decodability, not native output control.

The paired H2-minus-H1 change in categorical control is $0.993\,[0.972, 1.014]$. H1 remains near zero under its own linear classifier, while H2 is strongly readable. This weakens a global linear-head-mismatch explanation, but does not test nonlinear or distributed information at H1, nor whether decodable information governs native behavior. A secondary split that holds out complete worlds from every family gives $C = -0.014\,[-0.044, 0.019]$ at H1 and $C = 0.993\,[0.985, 0.998]$ at H2, with a change of $1.006\,[0.974, 1.036]$. Repair remains excluded from fitting.

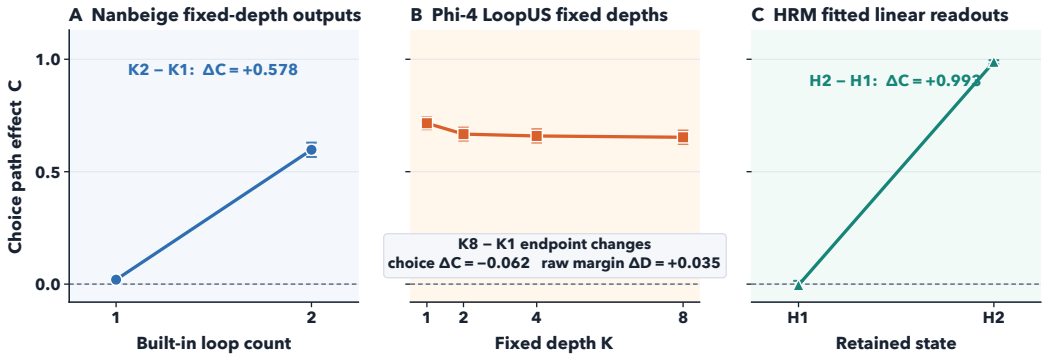

Figure 6: **Additional checkpoints show distinct depth profiles, while fitted readouts preserve HRM's H1-to-H2 separation. A:** Nanbeige's choice effect rises between its native fixed-depth one- and two-pass outputs; only K2 is the trained default endpoint. **B:** Across separate fixed-depth Phi-4 LoopUS runs, margin-control divergence appears as falling choice control and a growing raw-margin effect. **C:** Separately fitted linear classifiers give near-zero control at HRM's retained H1 and near-complete control at H2. Error bars show 95% bootstrap intervals; readout labels are comparable only within a system.

I.5 Ouro acquisition remains positive across the tested prompt formats

Ouro passed every condition on the 24-world competence set ($0.693\,[0.655, 0.730]$ at K4). On the separate 48-world intervention set, K1 remained near zero and all three planned lower bounds in Table 17 were positive.

Slice	D_4	$\Delta D_{4,1}$	$\Delta C_{4,1}$
Overall	$0.349\,[0.340, 0.359]$	$0.350\,[0.340, 0.360]$	$0.298\,[0.279, 0.317]$
Serialization			
Arrows	$0.349\,[0.330, 0.369]$	$0.345\,[0.326, 0.365]$	$0.346\,[0.315, 0.378]$
Labeled tuples	$0.395\,[0.367, 0.424]$	$0.401\,[0.373, 0.429]$	$0.323\,[0.276, 0.367]$
Compact JSON	$0.364\,[0.345, 0.382]$	$0.363\,[0.345, 0.382]$	$0.331\,[0.286, 0.375]$
Mapping sentences	$0.288\,[0.269, 0.306]$	$0.289\,[0.270, 0.308]$	$0.193\,[0.161, 0.224]$
Identifier morphology			
Uppercase, underscore	$0.390\,[0.367, 0.414]$	$0.394\,[0.370, 0.419]$	$0.365\,[0.326, 0.404]$
Lowercase, underscore	$0.339\,[0.316, 0.362]$	$0.336\,[0.313, 0.359]$	$0.151\,[0.112, 0.190]$
Title case, joined	$0.343\,[0.320, 0.366]$	$0.346\,[0.323, 0.368]$	$0.438\,[0.396, 0.479]$
Uppercase, joined	$0.324\,[0.307, 0.340]$	$0.322\,[0.305, 0.340]$	$0.240\,[0.206, 0.271]$
Question wording			
Wording 1	$0.385\,[0.360, 0.412]$	$0.392\,[0.366, 0.420]$	$0.331\,[0.289, 0.372]$
Wording 2	$0.359\,[0.340, 0.377]$	$0.362\,[0.343, 0.380]$	$0.341\,[0.307, 0.375]$
Wording 3	$0.333\,[0.317, 0.349]$	$0.327\,[0.311, 0.343]$	$0.109\,[0.083, 0.135]$
Wording 4	$0.319\,[0.295, 0.343]$	$0.317\,[0.292, 0.341]$	$0.411\,[0.359, 0.464]$

Table 17: **Ouro's native-exit acquisition remains positive across every tested format.** Brackets are 95% intervals. Rows below the overall result are coupled slices of the same 48 worlds, not independent tests.

Every slice is positive, so the effect does not depend on the original renderer. Node names are drawn from role-blind pools, so they do not reveal a node's role. The study is limited to these formats and this model and does not identify a discrete graph algorithm. Appendices E.2 and G give the design and analysis.

I.6 Balance and competence-set exclusion checks

Leaving out each relation family gives Ouro choice gains from 0.457 to 0.494 and positive raw gains. LoopUS choice changes remain near zero while all six raw gains remain positive. Across the 12 ordered terminal pairs, every Ouro choice and raw gain, and every LoopUS raw gain, is positive.

These summaries are descriptive. Removing the 48 competence worlds leaves the same depth profiles in the other 240. Ouro retains a choice gain of $0.484\,[0.453, 0.515]$ and raw gain of $0.280\,[0.265, 0.295]$ logits. LoopUS retains a choice change of $-0.002\,[-0.014, 0.009]$ and raw gain of $0.214\,[0.205, 0.224]$ logits.

J Three-edge path-control transitions

The main fictional experiment uses two-edge paths. Here we ask whether the same transition appears when the answer requires three relations. Ouro-2.6B exposes four native exits, K1–K4. One HRM-Text-1B forward pass retains H1 and produces H2; we apply the same released output head to both, although only H2 was trained to produce an answer. H1 is therefore a fixed-head intermediate readout, not a trained exit. Each model uses separate competence and intervention worlds, the same two-example prompt, and no weight updates. The independent unit is a graph world.

J.1 Selecting a competent three-edge interface

Competence required final-readout accuracy of at least 0.65, a 95% lower bound above 0.50, and accuracy of at least 0.55 in every queried-source and candidate-order arm. Direct prompting did not meet this threshold for Ouro. Adding one solved A example and one solved B example did, so we fixed that interface before evaluating separate 48-world competence and 144-world intervention sets. HRM used the same interface with its own separate sets. The lowest arm accuracy was 0.594 for Ouro and 0.708 for HRM. We evaluated the swap and repair only after every competence condition passed.

Split or interface	Worlds	Original-graph result and use
Ouro direct development	24	K4 accuracy $0.620\,[0.557, 0.677]$; below threshold
Ouro two-example development	same 24	K4 accuracy $0.714\,[0.667, 0.760]$; template selected
Ouro competence set	48	K4 accuracy $0.677\,[0.638, 0.716]$; competence passed
HRM interface development	24	H2 accuracy $0.776\,[0.714, 0.833]$; interface retained
HRM competence set	48	H2 accuracy $0.734\,[0.680, 0.792]$; competence passed

Table 18: **Both selected interfaces pass on the held-out competence sets.** Direct-prompt Ouro development is below threshold. Brackets are 95% intervals. The 24-world development estimates select or check the interface; the 48-world competence sets precede the separate 144-world intervention studies.

J.2 Three-edge task, intervention, and prompts

Each world contains four paths that share no nodes. The index $j \in \{1, 2, 3, 4\}$ identifies a path, and r_1, r_2, r_3 are its ordered relation labels. Source S_j reaches bridge B_j, second intermediate node M_j, and terminal T_j:

$$S_j \xrightarrow{r_1} B_j \xrightarrow{r_2} M_j \xrightarrow{r_3} T_j. \tag{5}$$

Paths 1 and 2 are queried; paths 3 and 4 are distractors. The focal terminals form an ordered pair from red, blue, green, and black. Six three-relation families crossed with all 12 ordered pairs and two replicates give $6 \times 12 \times 2 = 144$ intervention worlds, 24 per family. The separate 48-world

competence set is also balanced by family and terminal pair. Each world has a fixed shuffled record order.

Let G_O, G_S, and G_R denote the original, first-edge-swap, and repair graphs. In G_O, source S_1 reaches T_1 and source S_2 reaches T_2. Graph G_S exchanges the first edges, so each source reaches the competing terminal. Graph G_R retains that swap but exchanges the final terminal edges, restoring S_1 to T_1 and S_2 to T_2 after all three relations. All three graphs contain the same nodes, relations, and terminals. A rules-based solver verifies the answer supported by every prompt.

Each intervention world crosses three graph states, two queried sources, and two candidate orders: 12 target prompts per world and 1,728 per model. Ouro retains K1, K2, K3, and K4 from one native four-step run, yielding 6,912 scored rows. K1 versus K4 is planned; K2 and K3 are descriptive. One unmodified HRM forward pass retains both H1 and H2, yielding 3,456 rows. H2 is the native output; applying its fixed head to H1 gives the intermediate readout.

Target wording is identical across systems. Square brackets below mark fields filled from a world; they do not denote optional text. All other text is literal. The record line appears 12 times, once per shuffled edge.

```
This is a fictional directed graph.  Every node and record is
invented.  Use only the records below.

RECORDS
SUBJECT = [subject] | RELATION = [relation] | OBJECT = [object]
[11 further records in the fixed order]

QUESTION
Starting from [source], follow these relations in order:  [relation
1], then [relation 2], then [relation 3].  Which candidate terminal
code is reached?

CANDIDATES
A. [first terminal]
B. [second terminal]

Return exactly A or B and nothing else.
ANSWER =
```

The fixed system instruction is: "Use only the supplied fictional directed graph. Follow the named relations exactly. Your entire answer must be one label: A or B." Both systems precede the target with the same examples under `SOLVED THREE-EDGE EXAMPLE 1` and `SOLVED THREE-EDGE EXAMPLE 2`, then `TARGET PROBLEM`. Their nodes, relations, and terminal words differ from the evaluation sets. Ouro uses its released chat template; HRM uses the checkpoint's direct PrefixLM condition. A and B are each one token under both tokenizers. Demonstrations are in-context text only; weights do not change.

Readout mapping and inference. The estimands and decision rules from Appendix G apply without modification. The early and late readouts are K1 and K4 for Ouro, and H1 and H2 for HRM. Within each world, we average the two queried sources and two candidate orders. The recovery ratios $R_t^m = N_t/D_t$ and $R_t^c = Q_t/C_t$ divide the repair-minus-swap margin or choice increment by its original-minus-swap effect. One means full average recovery. State accuracy instead credits the terminal supported by each graph. The family-stratified bootstrap resamples whole worlds and keeps graph states, source arms, orders, and readouts paired.

J.3 THREE-EDGE RESULTS AND REPAIR

Table 19 reports every readout. Ouro is at chance at K1 and K2; a small K3 choice effect precedes the larger planned K4 effect. For HRM, the fixed output head shows a near-zero effect at H1 and a large effect at the native H2 output. This comparison identifies when control becomes readable through that head, not when the underlying computation begins. Margin magnitudes differ across models and should not be compared directly.

For Ouro, the planned K1-to-K4 margin gain is $0.348\,[0.323, 0.374]$, and the choice gain is $0.297\,[0.260, 0.335]$. At K4, the margin and choice repair increments are $0.323\,[0.285, 0.362]$ and

Readout	Original acc.	Swap acc.	Repair acc.
Ouro K1	$0.500\,[0.500, 0.500]$	$0.500\,[0.500, 0.500]$	$0.500\,[0.500, 0.500]$
Ouro K2	$0.500\,[0.500, 0.500]$	$0.500\,[0.500, 0.500]$	$0.500\,[0.500, 0.500]$
Ouro K3	$0.515\,[0.505, 0.525]$	$0.521\,[0.510, 0.532]$	$0.519\,[0.507, 0.531]$
Ouro K4	$0.659\,[0.635, 0.683]$	$0.638\,[0.613, 0.663]$	$0.627\,[0.602, 0.653]$
HRM H1	$0.500\,[0.500, 0.500]$	$0.500\,[0.500, 0.500]$	$0.500\,[0.500, 0.500]$
HRM H2	$0.739\,[0.707, 0.770]$	$0.753\,[0.722, 0.782]$	$0.746\,[0.716, 0.775]$

Readout	Margin effect D_t	Choice effect C_t
Ouro K1	$-0.001\,[-0.003, 0.001]$	$0.000\,[0.000, 0.000]$
Ouro K2	$0.004\,[0.000, 0.008]$	$0.000\,[0.000, 0.000]$
Ouro K3	$0.121\,[0.110, 0.133]$	$0.036\,[0.022, 0.051]$
Ouro K4	$0.347\,[0.322, 0.373]$	$0.297\,[0.260, 0.335]$
HRM H1	$0.000\,[-0.001, 0.001]$	$0.000\,[0.000, 0.000]$
HRM H2	$1.924\,[1.785, 2.065]$	$0.491\,[0.444, 0.537]$

Table 19: **Ouro acquires three-edge path control; HRM shows a fixed-head transition.** The upper panel reports graph-state accuracy; the lower panel reports margin and choice effects. Brackets are 95% intervals. K1 versus K4 and H1 versus H2 are planned; Ouro K2 and K3 are descriptive.

$0.265\,[0.222, 0.307]$, giving $R^c_{K4} = 0.892\,[0.776, 1.012]$. For HRM, the H1-to-H2 margin gain is $1.924\,[1.785, 2.065]$, and the choice gain is $0.491\,[0.444, 0.537]$. At H2, the repair increments are $1.962\,[1.775, 2.144]$ and $0.498\,[0.441, 0.554]$, giving $R^c_{H2} = 1.014\,[0.930, 1.101]$. Ouro's planned K1-to-K4 native-exit transition passes every acquisition and repair rule; HRM passes the corresponding retained-state transition and repair thresholds.

We also recomputed all four contrasts within relation family, ordered terminal pair, queried-source arm, and candidate order. Every partition estimate is positive for both systems. The smallest choice gain is 0.188 for Ouro and 0.375 for HRM; the smallest choice repair increment is 0.141 and 0.333. These slices are descriptive, not independent or adjusted tests.

J.4 Checks and scope

Independent examples. Development, competence, and intervention sets do not share identifiers or target prompts. Across Ouro's competence and intervention sets, all 2,304 nonterminal nodes are unique. No evaluation prompt or adapted prompt appears in a development set, the other model's set, or either solved example. The Ouro and HRM intervention sets share only the balanced terminal colors.

Paired readouts. For each prompt, the early and late readouts come from the same run. Ouro's readouts match both the public forced exits and independently truncated executions. On HRM, the released head applied to H2 exactly reproduces the checkpoint's stock 65,536-way logit vector. Applying that same fixed head to the H1 state gives the intermediate readout. Thus every readout contrast is paired within prompt and run.

Interpretation. For these checkpoints and fixed demonstrations, Ouro's native-exit acquisition extends from two-edge to three-edge paths. HRM shows a corresponding transition from its fixed-head H1 readout to its native H2 output. We do not test arbitrary path lengths, natural three-relation reasoning, open-vocabulary generation, Ouro branching, or a serial internal graph-traversal algorithm. Nor does this experiment separate recurrence from training and effective depth as the cause. Ouro exposes only four native readouts, so K4 marks the tested horizon and does not imply one loop per edge.

K Ouro's response follows the answer-bearing path

These structural tests apply to Ouro-2.6B. One places the same edit on or off the answer path; the other makes graph connectivity conflict with physical proximity. Together they ask whether the

fictional result can be reduced to a generic edit response or a nearest-record heuristic. After the base fictional competence screen passed, a separate 24-world development set checked execution of these prompt variants. The held-out intervention set contains 144 worlds, with two instances of each of six relation families and 12 ordered terminal pairs. A symbolic solver validates all 4,608 prompts. Appendix F.4 gives the prompts.

K.1 THE SAME EDIT MATTERS ON THE ANSWER PATH

Both conditions use the same first-hop transposition; only the queried component changes. The on-path response grows across exits, while the off-path response stays near zero (Table 20).

Raw logit effects

Exit	On path	Off path	Specificity
K1	$-0.001\,[-0.003, 0.002]$	$-0.002\,[-0.004, 0.000]$	$0.001\,[-0.001, 0.004]$
K2	$0.037\,[0.033, 0.041]$	$-0.002\,[-0.006, 0.002]$	$0.040\,[0.034, 0.045]$
K3	$0.195\,[0.182, 0.209]$	$-0.004\,[-0.009, 0.002]$	$0.199\,[0.183, 0.214]$
K4	$0.270\,[0.252, 0.291]$	$0.000\,[-0.006, 0.005]$	$0.271\,[0.251, 0.293]$

Exact-choice effects

Exit	On path	Off path	Specificity
K1	$0.000\,[0.000, 0.000]$	$0.000\,[0.000, 0.000]$	$0.000\,[0.000, 0.000]$
K2	$0.029\,[0.017, 0.041]$	$0.003\,[-0.006, 0.011]$	$0.026\,[0.011, 0.042]$
K3	$0.269\,[0.232, 0.307]$	$0.000\,[-0.017, 0.017]$	$0.269\,[0.229, 0.310]$
K4	$0.482\,[0.440, 0.523]$	$0.018\,[-0.004, 0.040]$	$0.464\,[0.415, 0.513]$

Table 20: **On-path versus off-path bridge transposition.** Raw effects are logits; choice effects are unitless; specificity is on minus off. Brackets are 95% intervals from 10,000 shared, relation-stratified whole-world bootstrap draws.

K4-minus-K1 specificity is $0.269\,[0.250, 0.291]$ logits and $0.464\,[0.415, 0.513]$ in choice. Both off-path K4 intervals contain zero, so this assay provides no evidence for a generic response to the transposition.

K.2 THE GRAPH STILL MATTERS WHEN NEARBY EVIDENCE POINTS ELSEWHERE

The locality test fixes the queried first hop and compares normal with transposed terminal bindings. In matched prompts, the terminal record nearest the queried bridge is also graph-connected to it. In crossed prompts, the nearest record supports the competing bridge, so a nearest-record strategy predicts the wrong answer.

The planned crossed K4-minus-K1 effects have positive intervals (Table 21). Local placement is associated with a larger response but does not determine its sign. Descriptively, all six relation-family estimates are positive for both planned effects on both scales. All 12 edited terminal-pair estimates are positive for raw specificity and the crossed-locality effect.

Locality remains a strong cue: at K2, matched and crossed raw effects are 0.116 and 0.014. At K3, the matched-minus-crossed difference of 0.424 logits exceeds twice the 0.164 crossed graph effect, although crossed accuracy reaches 0.614. The response remains aligned with connectivity under the adverse cue, while locality is associated with its magnitude and timing.

Raw logit effects

Exit	Matched	Crossed	Locality
K1	$-0.003\,[-0.005, 0.000]$	$-0.001\,[-0.003, 0.002]$	$-0.002\,[-0.006, 0.001]$
K2	$0.116\,[0.108, 0.124]$	$0.014\,[0.008, 0.019]$	$0.102\,[0.094, 0.111]$
K3	$0.587\,[0.557, 0.616]$	$0.164\,[0.147, 0.180]$	$0.424\,[0.401, 0.446]$
K4	$0.664\,[0.634, 0.694]$	$0.258\,[0.231, 0.285]$	$0.406\,[0.388, 0.424]$

Exact-choice effects and state accuracy

Exit	Matched choice	Crossed choice	Locality choice	Accuracy Matched / Crossed
K1	$0.000\,[0.000, 0.000]$	$0.000\,[0.000, 0.000]$	$0.000\,[0.000, 0.000]$	$0.500/0.500$
K2	$0.204\,[0.175, 0.234]$	$0.007\,[-0.009, 0.023]$	$0.197\,[0.165, 0.230]$	$0.602/0.503$
K3	$0.643\,[0.598, 0.689]$	$0.228\,[0.188, 0.270]$	$0.415\,[0.378, 0.451]$	$0.822/0.614$
K4	$0.740\,[0.694, 0.786]$	$0.438\,[0.378, 0.499]$	$0.301\,[0.260, 0.343]$	$0.870/0.719$

Table 21: **Graph-consistent control survives adverse locality.** Choice is the exact-choice effect and locality is matched minus crossed; brackets are 95% intervals. The planned crossed K4-minus-K1 effects are $0.259\,[0.232, 0.285]$ logits and $0.438\,[0.378, 0.499]$ in choice. Accuracy credits the graph-supported answer; accuracy entries are point estimates pooled over normal and transposed terminal bindings.

K.3 Secondary natural-language controls

These secondary controls use 2Wiki entities and relation phrases. Each eight-record prompt contains four two-edge chains, with the same first-hop transposition in the queried or disconnected pair. The crossed design makes each natural pair both focal and distractor. It contains 124 pairs in 62 blocks across 31 paths and includes both questions and candidate orders. Ouro uses native K1, K3, and K4; LoopUS uses separate fixed-depth K1 and K8 outputs. Categorical estimates use the first token where candidate answers differ and therefore depend on the answer boundary.

At Ouro K3 and LoopUS K8, the complete-answer likelihood effects are $5.251\,[4.646, 5.887]$ nats on path and $0.131\,[0.072, 0.193]$ off path; LoopUS gives $3.599\,[3.177, 4.068]$ and $0.089\,[0.037, 0.139]$. Their on-minus-off effects are $5.120\,[4.502, 5.766]$ and $3.510\,[3.078, 3.985]$ nats. These are not directly comparable to the primary four-record assay: this prompt has four rather than two chains and uses complete-answer log likelihood rather than a one-token A/B logit.

The secondary first-divergence analysis shows that the structural effect also changes the local candidate winner:

System	Readout	Specificity	Deep minus early
Ouro	K1	$0.002\,[-0.007, 0.012]$	
	K3	$0.516\,[0.448, 0.589]$	
	K4	$0.593\,[0.513, 0.672]$	$0.591\,[0.509, 0.673]$
LoopUS	K1	$0.316\,[0.246, 0.389]$	
	K8	$0.394\,[0.336, 0.455]$	$0.078\,[0.027, 0.127]$

Table 22: **Secondary natural on/off-path choice analysis.** Specificity is the on-minus-off first-divergence choice effect. Brackets are 95% intervals from 10,000 paired path-level draws: 31 paths for Ouro and 30 for LoopUS after the exclusion below. Deep minus early is K4 minus K1 for Ouro and K8 minus K1 for LoopUS.

The natural locality design crosses record position, matched or crossed, with terminal mapping, normal or transposed. The first hops are already swapped. Within each position condition, the graph effect is the transposed-minus-normal choice contrast. Thus, graph connectivity and the nearest terminal record either agree or conflict. Across 124 whole-pair units and 31 relation paths, Ouro's first-divergence graph effect is $0.004\,[-0.012, 0.020]$ at K1 and $0.373\,[0.282, 0.467]$ at K3 when locality agrees. Under adverse locality it is $0.007\,[-0.004, 0.020]$ at K1 and $0.211\,[0.145, 0.278]$ at K3, a gain of $0.204\,[0.136, 0.272]$. Local placement amplifies the K3 choice effect by $0.162\,[0.112, 0.215]$

without determining its sign. Complete-answer K3 effects are $2.874\ [2.312, 3.425]$ nats when matched and $1.662\ [1.277, 2.050]$ under adverse locality.

Inference follows Appendix G.3. Categorical scoring uses the leading-space boundary for Ouro and bare assistant-boundary tokens for LoopUS. One LoopUS path is ineligible for first-divergence scoring because `poison` is a strict prefix of `poisoning` and the saved scores omit the end-of-sequence token. We therefore exclude it only from the categorical analysis; the complete-answer margin retains all 31 paths.

L What moves path control earlier or later

What makes path control appear at an early readout in some systems and only later in others? A controlled experiment varies parameter tying and which readouts receive supervision. Separate linear classifiers test whether the terminal output head misses early information. An ordinary untied Transformer provides a feed-forward comparison, and a training-free recurrent wrapper tests whether the effect requires recurrent training.

L.1 Controlled supervision experiment

The graph-memory study in Appendix B begins with a two-by-two comparison: tied versus untied blocks, each trained with K4-only or K1–K4 supervision. Inputs have four shuffled two-edge paths that share no nodes. Tied models apply one block four times; untied models apply four initially identical blocks once. All cells therefore use four attention and feed-forward applications, although parameter counts differ. Each condition has eight runs and 600 held-out worlds per run; intervals resample runs and paired worlds.

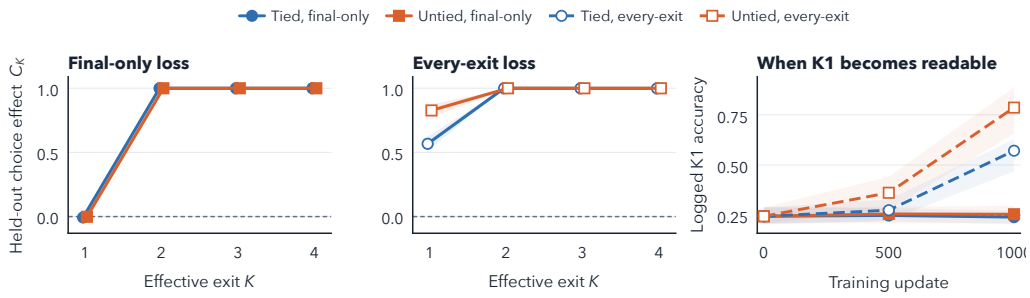

Figure 7: **Supervising every readout makes path control appear earlier in both tied and untied models.** The left and middle panels show graph-consistent choice control at each readout. Lines average eight seeds; bands give descriptive 95% intervals over seed means. The right panel shows K1 training accuracy, with the dotted line at four-way chance.

Table 23: **Supervising every readout makes control appear earlier.** Each entry gives the K1-to-K4 exact-choice gain and its 95% bootstrap interval. The final column is the reduction in that gain when every readout is trained. Intervals resample eight independent training runs and their held-out worlds.

Blocks	Final readout only	Every readout	Earlier-control shift
Tied	$1.007\ [0.983, 1.032]$	$0.432\ [0.376, 0.490]$	$0.575\ [0.504, 0.639]$
Untied	$1.002\ [0.977, 1.026]$	$0.173\ [0.125, 0.227]$	$0.829\ [0.761, 0.891]$

All 32 runs reach at least 0.9983 K4 accuracy. Under K4-only supervision, the tied-minus-untied difference in K1-to-K4 choice gain is $0.005\ [-0.015, 0.024]$. Supervising every readout reduces this late gain more for untied blocks: the difference-in-differences is $0.254\ [0.198, 0.311]$, with the same ordering in all eight paired seeds. Final control and accuracy remain near one, so supervision changes when control appears rather than final competence.

Intermediate supervision schedules. We next add K3–K4 and K2–K4 supervision, shown in Figure 4a. Supervised readouts receive equal weight: the K4-only, K3–K4, K2–K4, and K1–K4 schedules assign weights 1, $1/2$, $1/3$, and $1/4$ to each included loss. Total loss weight is therefore one in every condition. The model, data order, optimizer, 1,000 updates, and 600 held-out worlds are unchanged. Each new architecture-by-schedule cell uses the same eight seeds. This adds 32 runs to the original 32. Table 24 reports the K1 effects.

Blocks	Supervise K3–K4	Supervise K2–K4
Tied	$0.004\,[-0.019, 0.026]$	$-0.003\,[-0.022, 0.015]$
Untied	$0.003\,[-0.021, 0.027]$	$0.004\,[-0.022, 0.029]$

Table 24: **The two intermediate schedules leave the K1 choice effect near zero.** Each entry is the K1 choice path effect C_1 and its 95% interval. Intervals use 10,000 hierarchical draws over eight seeds and paired evaluation worlds.

For tied blocks, the mean C_1 across the two intermediate schedules differs from K4-only by $0.007\,[-0.007, 0.022]$; for untied blocks, the difference is $0.005\,[-0.011, 0.021]$. Including K1 in the supervision schedule instead raises C_1 to $0.567\,[0.511, 0.623]$ and $0.827\,[0.773, 0.876]$, respectively. All 64 runs pass the K4 competence check; the lowest seed-level accuracy is 0.9983. Under these equal-total-weight schedules, supervision beginning at K2 or K3 does not make choice control visible at K1. Because adding a supervised readout also redistributes weight among the later losses, this experiment does not isolate the earliest supervised readout from the loss weighting.

L.2 Depth-specific readouts

Could the terminal head miss an answer already present at K1? Because the original weights were not retained, we retrained both final-only conditions. At each depth, we fit a separate linear classifier with a bias term to 12,000 new examples, after standardizing its inputs. Each classifier uses an L_2 penalty of 1 and is tested on 600 paired worlds per seed. The shared terminal head reproduces the original depth profile.

Graph-consistent choice control Architecture and readout	K1 C [95% CI]	K2 C [95% CI]
Tied, shared terminal head	$-0.004\,[-0.029, 0.019]$	$1.000\,[0.999, 1.000]$
Tied, depth-specific classifier	$-0.007\,[-0.026, 0.014]$	$1.000\,[1.000, 1.000]$
Untied, shared terminal head	$0.001\,[-0.022, 0.022]$	$1.000\,[1.000, 1.000]$
Untied, depth-specific classifier	$-0.014\,[-0.039, 0.009]$	$1.000\,[1.000, 1.000]$
State accuracy Architecture and readout	K1 accuracy [95% CI]	K4 accuracy [95% CI]
Tied, shared terminal head	$0.249\,[0.235, 0.262]$	$1.000\,[1.000, 1.000]$
Tied, depth-specific classifier	$0.243\,[0.235, 0.251]$	$1.000\,[1.000, 1.000]$
Untied, shared terminal head	$0.253\,[0.238, 0.266]$	$1.000\,[1.000, 1.000]$
Untied, depth-specific classifier	$0.244\,[0.234, 0.253]$	$1.000\,[1.000, 1.000]$

Table 25: **Depth-specific classifiers do not reveal linearly readable K1 control.** The upper panel reports graph-consistent choice control C at K1 and K2; the lower panel reports state accuracy at K1 and K4. Intervals use 10,000 hierarchical draws over eight seeds and paired worlds.

At K1, state accuracy remains at four-way chance and choice control remains near zero under both readouts. Both measures are essentially perfect by K2 and remain so at K4. The depth-specific classifiers give K4-minus-K1 choice gains of $1.007\,[0.987, 1.027]$ for tied blocks and $1.014\,[0.991, 1.038]$ for untied blocks. The transition therefore does not arise only because the shared terminal head is applied to an early state. This linear analysis of a symbolic four-block model does not rule out nonlinear information at K1 or explain timing differences among pretrained models.

L.3 A STANDARD UNTIED TRANSFORMER

We project each of Qwen3-4B-Instruct-2507's 36 untied post-layer states through its released final normalization layer (RMSNorm) and output layer (Qwen Team, 2025). The 288-world panel has 48 worlds per relation family and pairs graph state, queried source, and candidate order. Intervals use 10,000 relation-stratified whole-world draws.

In Figure 4, L1–L35 are fixed-head intermediate-layer readouts; only L36 was trained for output through this head. Choice control is zero at L1. L24 is the first layer with a positive C lower bound and above-chance original and swap accuracy: $C = 0.054\,[0.039, 0.069]$. At native L36, $C = 0.582\,[0.553, 0.610]$; original, swap, and repair accuracies are $0.775\,[0.753, 0.796]$, $0.807\,[0.790, 0.823]$, and $0.814\,[0.797, 0.830]$. Repair recovers $1.070\,[1.039, 1.101]$ of the final margin contrast; L36 logits match ordinary output exactly.

This nonmonotonic curve is not a serial algorithm trace. It shows that path control can also become visible late under a fixed head across ordinary untied layers. The controlled graph-memory task likewise shows delayed control with untied blocks, so parameter reuse is not necessary for the phenomenon in that task. Feed-forward layers nevertheless differ from recurrent inference-time loops.

We also fit a separate linear classifier with bias at each layer. Under limits of 100, 200, and 1,000 optimization steps, L18 is the first layer whose choice effect has a positive lower bound and whose original and swap accuracies both exceed chance. Every fit uses its full step budget, however, and final-layer choice control changes from 0.834 to 0.819 to 0.772 across those budgets. These fitted-classifier results are therefore exploratory; the fixed-head curve is the primary analysis.

Training-free recurrent wrapper. The wrapper in Appendix B uses the same 288-world two-edge panel. A 48-world original-state subset passes the K1 competence rule (overall accuracy at least 0.65, lower bound above 0.50, and each arm and order at least 0.55): $0.760\,[0.708, 0.812]$ overall and $0.771/0.750$ for both the two arms and two orders. The depth analysis uses all 288 worlds. Its 10,000 draws resample 48 worlds within each of six relation families while keeping depths, graph states, arms, and orders paired. From K1 to K6, choice control grows by $0.029\,[0.010, 0.048]$ and the raw-margin response by $1.003\,[0.890, 1.119]$ nats. Each K is a separate run with a different number of wrapper updates, so this is a depth profile rather than one recurrent trajectory.

M FROM FIXED-CHOICE CONTROL TO GENERATED ANSWERS

The main assay compares two predefined answers. We first remove the displayed answer list while retaining fixed-continuation scoring. We then perturb hidden states at the answer positions and test free generation with no supplied continuation. A final analysis varies the token boundary used to score each candidate.

Free-generation decision rule. We count generation as successful only if every condition below holds. The later graph effect must be at least 0.15, the depth gain at least 0.10, and both 95% lower bounds must exceed zero. Pooled later accuracy must have a lower bound above 0.50, with point accuracy above 0.50 in each graph. Both effects must also be positive in each query arm and, for LoopUS, each candidate order. Ouro requires the normalized complete output to equal the graph-supported answer. The LoopUS parser, fixed before generation, also accepts the wrapper `ANSWER = {candidate}`. We report literal string matching only as a descriptive check.

M.1 MOST PATH SENSITIVITY SURVIVES WITHOUT DISPLAYED OPTIONS

LIST displays both answers in counterbalanced order; NO_LIST removes only that block. The evidence, question, answer cue, K1/K3 readouts, and teacher-forced continuations are unchanged. Each answer token is scored after its preceding answer tokens. The same neutral prompt and analysis cover all 124 2Wiki pairs, without competence or outcome filtering. Brackets below mark substituted fields.

Test	Effect and gain or control	Accuracy / interpretation
Displayed-answer removal; effects in nats		
LIST	$D_3 = 0.703\,[0.539, 0.872]$; $G_{31} = 0.701\,[0.538, 0.874]$	Baseline
NO_LIST	$D_3 = 0.622\,[0.449, 0.813]$; $G_{31} = 0.632\,[0.447, 0.829]$	Retains 88.5% effect; 90.1% gain
State access; recovery ratio		
Candidate-state transplant	$0.222\,[0.147, 0.301]$; mirrored $-0.087\,[-0.149, -0.043]$	Partial recovery
Nearby-state controls	Marker $0.006\,[-0.001, 0.018]$; post-object $0.003\,[-0.006, 0.019]$	Little recovery
Greedy generation; emitted-identity effects		
Ouro 2Wiki NO_LIST (K3)	Effect and gain $0.198\,[0.115, 0.288]$	Accuracy $0.540\,[0.492, 0.591]$; accuracy check fails
Ouro MuSiQue answer-only (K3)	Effect and gain $0.142\,[0.082, 0.205]$	Accuracy $0.503\,[0.461, 0.545]$; check fails
LoopUS 2Wiki listed (K8)	Effect $0.086\,[0.051, 0.124]$; gain $-0.085\,[-0.152, -0.024]$	Accuracy $0.129\,[0.085, 0.175]$; check fails

Table 26: **Option removal preserves most likelihood control, candidate states carry part of that preference, and free generation fails the stated criteria.** The two planned localization comparisons use 97.5% intervals; all others use 95%.

```
Using only the evidence below, follow the two linked triples that begin
from the entity in the question.
EVIDENCE [four SUBJECT = ...  | RELATION = ...  | OBJECT = ...  records]
QUESTION [natural question]
[LIST only: CANDIDATE ANSWERS, followed by the two answers]
Return exactly the final OBJECT string reached by the path and nothing
else.  Do not number or explain it.  ANSWER =
```

All other rendered content is identical. Let z^I_{iK} be the complete-continuation reference-minus-other margin from Appendix F.5, with the same leading-space convention for $I \in \{\text{LIST}, \text{NO_LIST}\}$. Each pair supplies two graph states: O supports the reference answer, and S supports its paired answer. After averaging the query arms and, for LIST, the two answer orders, define $D_{I,K} = \frac{1}{2}\mathbb{E}_i[\bar{z}_{i,O,I,K} - \bar{z}_{i,S,I,K}]$. The factor $1/2$ converts the symmetric difference between opposite graph orientations to the average shift per orientation. This interface study uses only these two states. The depth gain is $G_{I,31} = D_{I,3} - D_{I,1}$. Intervals use 10,000 shared draws that resample 31 relation paths and then four whole pairs within each path.

NO_LIST retains $88.5\%\,[75.9, 100.9]$ of the K3 effect and $90.1\%\,[74.6, 105.9]$ of the gain. LIST-minus-NO_LIST differences are $0.081\,[-0.006, 0.166]$ and $0.069\,[-0.041, 0.171]$ nats. This does not show equivalence: the answers still occur in the evidence, and the endpoint still scores fixed continuations. The ratios use the paired LIST baselines in Table 26.

M.2 Candidate positions carry part of the preference

This intervention asks where the output preference can be changed without altering the output interface. Donor and recipient graphs support opposite answers. Before K3, we replace Ouro-2.6B's two 2,048-dimensional pass-2 states at the unchanged one-token answer mentions with the donor states, while holding every other state fixed. Position controls instead replace the preceding number-marker states or unchanged post-object evidence states. A mirrored control moves the candidate states the same L_2 distance away from the donor. Replacing the full state or final readout gives exact positive controls.

We exclude the 24 discovery pairs and require one-token answers among the remaining 100. Let a_1 and a_2 be the K3 margins toward the supported answer in the two graph states of a pair, for one query and answer order. Inclusion requires $a_1 > 0$, $a_2 > 0$, and $a_1 + a_2 \geq 1.875$ nats, so both graphs must favor their supported answer and the two opposing margins must sum to at least 1.875 nats. This leaves 51 units from 29 pairs and 18 relation paths, or 102 directional transplants.

Let o_R, o_D, and o_P be the log odds in favor of the donor's answer in the recipient, donor, and patched runs. The unclipped recovery is $\rho = (o_P - o_R)/(o_D - o_R)$. Its denominator equals $a_1 + a_2$ and is therefore at least 1.875. We average arms, orders, and directions within each pair and report the median over 29 equally weighted pairs. The 10,000-draw bootstrap resamples relation paths and then pairs. Candidate mentions and post-object evidence states are the two planned localization tests. These tests and their controls use 97.5% intervals; all other transplant intervals use 95%.

Candidate-state recovery is positive in all 29 pairs; its median separation from the mirrored control is $0.306\,[0.238, 0.438]$. Nearby controls recover little, while full-state and final-readout controls recover exactly one. Candidate positions can therefore carry part of the output preference when answer mentions are present. This does not show that these positions are necessary, unique, or where path composition occurs.

M.3 Likelihood control does not guarantee competent generation

Greedy generation starts at the answer boundary with no supplied continuation. It takes the full-vocabulary argmax for at most 24 tokens, without sampling or length penalty, and stops at EOS, chat-turn end, or the first line break. The parsers follow the rules above. Invalid or ambiguous outputs receive zero for both signed identity credit and accuracy. Valid outputs receive $+1$ for the original identity and -1 for its pair. The signed effect measures the shift under the graph edit; accuracy requires a fully correct output. These rules were fixed before generation.

On 2Wiki, Ouro's effect and gain meet their thresholds, but the accuracy check fails. On MuSiQue, the gain exceeds 0.10, while the K3 effect misses 0.15 by 0.008 and the accuracy check fails. At K3, 438/496 2Wiki outputs and 328/380 MuSiQue outputs match a candidate. LoopUS has a sub-threshold K8 effect, negative gain, and 128/744 candidate-matching K8 outputs. Two additional LoopUS scoring rules accept either a displayed list label or a unique candidate mention. Because these rules were defined after inspecting the outputs, we report them only as sensitivity analyses. Path control over a defined comparison therefore does not guarantee reliable free-form generation.

M.4 Scoring-boundary sensitivity

We also repeat the analysis at the token boundary used by generation. Primary scoring begins with a leading space, whereas generation begins directly at the assistant boundary. The bare score replays the first token without that space and teacher-forces the remainder. The combined score is the log-sum-exp of the complete leading-space and bare scores. Here D_K is the original-minus-swap margin response and C_K is its sign-based choice counterpart, with half credit for equal scores. These are sensitivity analyses; the primary endpoints remain unchanged. K3 pair-choice accuracy is $(1+C_3)/2$: the fraction of graph-state choices favoring the supported candidate, with half credit for ties.

Dataset	Score	D_1	D_3	$\Delta D_{3,1}$
2Wiki	Bare	$0.008\,[-0.037, 0.051]$	$0.495\,[0.354, 0.647]$	$0.487\,[0.330, 0.650]$
	Combined	$0.010\,[-0.032, 0.049]$	$0.496\,[0.355, 0.649]$	$0.486\,[0.332, 0.646]$
MuSiQue	Bare	$0.021\,[-0.017, 0.066]$	$0.378\,[0.258, 0.502]$	$0.357\,[0.229, 0.489]$
	Combined	$0.019\,[-0.018, 0.061]$	$0.378\,[0.259, 0.502]$	$0.359\,[0.235, 0.490]$

Dataset	Score	C_1	C_3	$\Delta C_{3,1}$	K3 pair-choice accuracy
2Wiki	Bare	$0.000\,[-0.032, 0.032]$	$0.214\,[0.117, 0.319]$	$0.214\,[0.105, 0.331]$	$0.607\,[0.558, 0.659]$
	Combined	$0.000\,[-0.028, 0.028]$	$0.206\,[0.109, 0.306]$	$0.206\,[0.101, 0.319]$	$0.603\,[0.554, 0.653]$
MuSiQue	Bare	$-0.011\,[-0.032, 0.011]$	$0.163\,[0.100, 0.232]$	$0.174\,[0.111, 0.237]$	$0.582\,[0.550, 0.616]$
	Combined	$-0.005\,[-0.026, 0.016]$	$0.184\,[0.116, 0.253]$	$0.189\,[0.121, 0.258]$	$0.592\,[0.558, 0.626]$

Table 27: **Ouro's K1-to-K3 choice gain remains positive when candidate scoring begins at the generation boundary.** All intervals are 95%; equal candidate scores receive half credit. The panels contain 496 rows from 124 2Wiki pairs and 380 rows from 95 MuSiQue pairs, using the same relation-path clustering as the primary analysis and 10,000 draws.

Among candidate-matching K3 outputs with unequal candidate scores, bare-token scores select the emitted answer in 98.4% of 2Wiki rows and 98.5% of MuSiQue rows. Combined scores agree on 98.4% and 97.9%, compared with 66.5% and 66.7% for leading-space scores. Relative to leading-space scoring, the bare boundary reverses the preferred candidate on 34.4% $[26.9, 42.4]$ and 35.4% $[28.7, 42.2]$ of rows. Ouro's K1-to-K3 choice gain remains positive under both alternatives (Table 27).

Taking the better rank across the leading-space and bare spellings, the correct candidate's median rank moves from 1,723 at K1 (interquartile range 143–8,167) to 1 at K3 (1–2) on 2Wiki. On MuSiQue, it moves from 1,135 (93–6,997) to 1 (1–3). The correct candidate is in the K3 top ten for 492/496 2Wiki rows and 370/380 MuSiQue rows.

Dataset	Correct prefix	Wrong prefix	Shared prefix	Outside both prefixes
2Wiki ($n = 496$)	263 (53.0%)	165 (33.3%)	8 (1.6%)	60 (12.1%)
MuSiQue ($n = 380$)	189 (49.7%)	136 (35.8%)	4 (1.1%)	51 (13.4%)

Table 28: **Most generation errors begin with the wrong candidate or outside both candidate prefix sets.** Greedy K3 first tokens are classified using leading-space and bare-token prefixes. Correct-prefix generations finish correctly on 262/263 2Wiki and 188/189 MuSiQue rows; every wrong-prefix generation finishes with the wrong candidate.

Outside-prefix starts produce invalid outputs on 57/60 2Wiki rows and all 51 MuSiQue rows, with three 2Wiki recoveries. Shared prefixes split between the two answers.

Prompt-only replay matches 98.6% of stored 2Wiki K3 first tokens and 98.2% of MuSiQue tokens; K1 agreement is exact. Analyses use stored generations because small numerical differences can change greedy output. Token-boundary choice explains much of the likelihood-to-generation mismatch, but free generation still fails the original accuracy criterion.

N WHERE THE ASSAY IS NOT INFORMATIVE

Table 29 lists settings where the model does not solve the unchanged task, so we do not estimate swap or repair effects. A failed screen makes the assay uninformative for that interface; it is not evidence of a zero intervention effect.

Interface specificity. These fictional-graph failures concern only that grammar: LoopUS-1.7B passes the natural panels (Appendix H). On natural 2Wiki, Ouro-1.4B has $D_4 = 4.037\,[3.481, 4.632]$ nats, $D_4 - D_3 = 2.124\,[1.767, 2.527]$, and $R_4^m = 0.658\,[0.563, 0.742]$. This second checkpoint shows a similar late increase, but is neither a held-out replication nor evidence of a scale effect.

Branching and stack reuse. Ouro-2.6B misses every required K4 subgroup threshold on the branching task. The author-released AnchorFix seed-42 pair compares 16 distinct blocks with one eight-block stack applied twice (Lee et al., 2026; Lee, 2026a;b). Its `weight_tying=false` setting refers to tying the input embeddings to the output head, not to block reuse. The branching screen and both checkpoints remain at chance, so this pair cannot compare stack reuse.

Longer paths. On its direct three-edge development panel, Ouro-2.6B has choice-control gain $0.214\,[0.141, 0.281]$ but misses the 0.65 final-accuracy threshold. The direct LoopUS ladder uses separate 24-world panels at three, four, and five edges, each with four disjoint paths and fresh opaque nodes. Only K1 and K8 are scored; Table 29 reports K8 accuracy. A different 24-world screen reuses the two solved three-edge examples from the Ouro and HRM studies. Its K1/K8 accuracies are $0.495/0.516$; K8 arm accuracies are $0.542/0.490$ and order accuracies $0.521/0.510$. Because this screen failed, we did not run the planned 48-world check or 144-world intervention panel. We retained its prompts and outcomes, but not the earlier planning note. Neither screen supports a conclusion about acquisition, sharpening, margin-control divergence, repair, or architecture. Appendix J reports the passing three-edge tests; four- and five-edge paths and natural three-relation questions remain unconfirmed.

Huginn. Separate, nonoverlapping 2Wiki K16 and K32 panels each contain 248 rows from 31 paths, two query arms, two candidate orders, and two initial-state seeds; within a seed, all variants of a pair share the same initialization. K16 slice accuracies span 0.472–0.492, and K32 arm accuracies

are $0.597/0.452$. Both fail the 0.60 overall, 0.50 lower-bound, and 0.55 subgroup rules. Since the panels differ, we do not estimate a paired gain. We verified the checkpoint identity, and a manual K4 execution reproduced the native output; we did not independently verify its training (Geiping et al., 2025; TOMG Group, University of Maryland, 2025).

DiscoLoop. As of 22 September 2026, the paper page and its linked resources provided no author code or runnable 440M checkpoint (Fu et al., 2026). We found one similarly named Hugging Face artifact that was not linked from the paper and declares a 1,936-layer Qwen configuration (DiscoLoop, 2026).

Screen	Accuracy [95% CI]
Ouro-1.4B, fictional 2-edge, K4	0.570 $[0.529, 0.612]$
LoopUS-1.7B, fictional 2-edge, K8	0.529 $[0.490, 0.568]$
Ouro-2.6B, branching development, K4	0.516 $[0.503, 0.531]$
Ouro-2.6B, direct 3-edge development, K4	0.620 $[0.557, 0.677]$
LoopUS-8B, direct 3/4/5-edge, K8 (24 worlds each)	0.521 $[0.479, 0.562]$ / 0.505 $[0.469, 0.542]$ / 0.490 $[0.469, 0.500]$
LoopUS-8B, shared-demo 3-edge, K1/K8	0.495 $[0.453, 0.536]$ / 0.516 $[0.474, 0.557]$
Huginn 2Wiki, separate K16/K32 panels	0.482 $[0.456, 0.504]$ / 0.524 $[0.500, 0.549]$
AnchorFix base/looped, fictional 2-edge	0.500 $[0.484, 0.516]$ / 0.500 $[0.500, 0.500]$

Table 29: **These settings do not support an intervention estimate.** Fictional thresholds follow Appendix G.2. For Huginn, overall accuracy must be at least 0.60, its lower bound must exceed 0.50, and each arm, order, and seed accuracy must be at least 0.55.