When Does Decodable Task Information Constitute Agent State? Dissociating Information, Currentness, and Behavioral Use in Live Coding Agents

Anonymous authors
Paper under double-blind review

Abstract

Probing is widely used to study task-relevant information in a model's internal representations, including information about the state of an ongoing task. For an acting agent, however, treating such information as state raises two further questions. Does it track the task as it changes, and does it shape the actions the agent takes? We study these questions in coding agents working on real software repositories, recording model activations and using task variables whose values can be determined directly from the environment. To separate current state from recency, we create conflicts in which an observation from the environment provides the current value, while stale evidence presented later asserts a different one. Our results empirically dissociate decodable information, authoritative currentness, and behavioral use. In our primary setting, where the prompt explicitly identifies the authoritative source, the current task value remains strongly decodable and continues to guide the agent's natural action. Yet a readout trained on ordinary cases fails at the frozen decision threshold when stale evidence is presented last. When current and stale values vary independently during fitting, however, the resulting readout transfers to held-out conflict data in every eligible case we test. In a separate construction, authority must be inferred from provenance rather than explicit labels, and we vary presentation order. The readout of the current value remains strong regardless of whether current or stale evidence appears first. We complement these measurements with causal interventions that separate changes in an internal score, in the decision process, in the emitted action, and in the final action. Effects depend on where the intervention enters the computation, and an action changed at one decision point may still be restored later in the trajectory. Together, these results show that evidence for agent state extends beyond representation to how task information enters decisions and persists as the task evolves.

1 Introduction

Probing has become a common way to study what information language models encode. Properties of modeled worlds, entities, and programs can often be decoded from hidden activations, including properties of the code being manipulated during repository-level tasks (Li et al., 2023; Gurnee & Tegmark, 2024; Li et al., 2021; Burns et al., 2023; Silva et al., 2026). For a live coding agent, however, the problem is different. Its task changes as it reads files, edits code, runs tests, and observes the results. A useful notion of state should therefore reflect what is currently true about the task and guide the agent's subsequent actions. Decodability alone does not provide sufficient evidence of such state. Task information may remain accessible after it becomes stale or without influencing the agent's behavior.

The problem becomes clear when current and stale evidence coexist. Consider a file whose current contents are X, established by a truthful read, while later stale evidence asserts Y. Retaining both pieces of information is legitimate. The question is whether the agent's internal representation continues to track the authoritative value X rather than the stale value Y. Even if it does, this does not establish that the authoritative value guides the agent's next action. We therefore separate three

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.

questions that are often conflated when discussing agent state. These are *decodable information*, *authoritative currentness*, and *behavioral use*. The three measurements support different claims and require different evidence.

We measure them separately in live coding agents. Hidden activations are captured online from the acting model during its own trajectories, while the current value of each task-state variable is determined directly from the environment rather than from human or model annotations. In ordinary trajectories, however, the authoritative value is often also the most recently presented one, so a successful readout could simply be following recency. To separate the two, we construct cases in which an observation of the current environment establishes the authoritative value, while stale evidence asserting a different value is placed later in the agent's context. Decodability is evaluated from the hidden activations, and behavioral use from the agent's natural action. For currentness, we train a readout on cases where the authoritative value is also the most recently presented value, and test whether it continues to track that value under conflict. We apply this framework to seven task-state families across six models.

The results reveal a clear dissociation among these measurements. In the primary file-content repair setting, task information is strongly decodable, while an agreement-trained readout fails the frozen currentness test when stale evidence is presented last. Separating current and stale values during fitting recovers a readout that transfers to held-out conflict data. The recovery remains strong in a separate factorial that varies presentation order and removes explicit authority labels. Natural behavior nevertheless follows the authoritative value on 98.9% of the primary conflict cases. Across six models and seven state families, currentness and behavioral use vary while task state remains decodable. These results show that decodability, authoritative currentness, and behavioral use provide different evidence about agent state.

We complement these measurements with causal interventions. In the primary file-content repair setting, replacing an internal activation with one from the opposite task condition can shift the generated action toward the opposite choice, yet the model can later revisit the decision and restore its original action. A complementary requirement-satisfaction experiment shows why timing matters. A calibrated intervention can change an internal decision score without changing behavior when applied before the model chooses between the candidate actions, while the behavioral effect appears at the point where that choice is made and replicates in a pre-specified held-out sample. Changing an internal quantity, changing an action during generation, and changing the final action are therefore not equivalent outcomes.

Contributions. (1) We introduce an authority–recency measurement framework that separates decodable information, authoritative currentness, and behavioral use in live coding agents. (2) Separating current and stale values during fitting recovers a readout of the current value for every eligible transport failure we test. A factorial that varies presentation order and removes explicit authority labels shows that the recovery is not explained by a fixed block position. Across models and families, currentness and behavioral use vary while task state remains decodable. (3) Causal interventions further separate decodability, influence on the decision process, emitted action, and persistence to the final action. Interventions can change the generated action when they reach the decision process, while earlier interventions may alter an internal score without changing behavior, and an immediate action change can later be revised when the model revisits the decision.

2 Related Work

Latent representations and state tracking. A large literature uses probes to study structured information in language-model activations. Linear readouts recover properties of game states, spatial and temporal structure, semantic variables, and other structured features represented in model activations (Li et al., 2023; Gurnee & Tegmark, 2024; Li et al., 2021; Burns et al., 2023). Propositional probes have further recovered structured world-state information even when a model's output is unfaithful to its context (Feng et al., 2025). More directly, Li et al. (2025) study how language models represent and update evolving state, identifying distinct internal mechanisms for state tracking. In coding agents, Silva et al. (2026) show that hidden activations encode properties of the evolving program and can predict outcomes of future edits.

Probe success, however, does not by itself establish how decoded information is used. Work on probe interpretation has long distinguished information accessibility from functional relevance (Hewitt & Liang, 2019; Belinkov, 2022; Elazar et al., 2021). Recent results make this distinction particularly explicit. Sudheendra & Srivastava (2026) show that logical-validity information can remain strongly decodable while behavior does not reliably express it, and that interventions along the corresponding probe direction can have weak or nonspecific effects. That work, like the earlier probing literature, separates what is accessible from what is expressed or causally relevant. It does not ask which value a readout tracks when the underlying task state evolves. In ordinary trajectories the authoritative current value is also the most recently presented one, so a readout that appears to track current state may instead reflect recency. Our conflict test reverses the presentation order of current and stale evidence to separate authoritative currentness from availability and behavioral use. Our interventions then distinguish immediate causal effects from effects that persist through later decisions.

Conflicting evidence and recency. Language models have also been studied under conflicts between contextual evidence and parametric knowledge, and between incompatible pieces of context (Longpre et al., 2021; Xie et al., 2024). The position of evidence can strongly affect how information is used in long contexts (Liu et al., 2024). These results show that both conflict and recency shape model behavior, and they expose a measurement problem for evolving task state. Determining whether an internal readout tracks the authoritative current value requires authority and recency to make different predictions.

Dynamic state and agent memory. State tracking is also a central problem for long-running agents. Prior systems maintain explicit summaries or memories across interactions, retrieve earlier experience, or track task state to support subsequent decisions (Packer et al., 2023; Park et al., 2023; Shinn et al., 2023; Rozanov & Rei, 2025). More recent work focuses directly on information that becomes outdated. STALE evaluates whether agents recognize that earlier memories are no longer valid and whether updated information carries through to downstream behavior (Chao et al., 2026). Supersede studies the gap between storing a revised fact and using its current value (Patel, 2026), while StateMemBench evaluates whether memory systems answer from current rather than superseded state over long interactions (Fan et al., 2026). This work makes current-versus-superseded state a first-class problem at the level of memory contents and behavioral outputs. It leaves open the complementary question of how that distinction is represented inside the acting model, especially when both current and superseded information remain available.

Causal interventions and self-repair. Causal interventions provide stronger evidence than passive decoding by directly changing internal activations. Activation patching, representation erasure, and steering have been used to test whether identified representations contribute to model outputs (Elazar et al., 2021; Zhang & Nanda, 2024; Hase et al., 2023). At the same time, causal localization and downstream behavioral effects can disagree, and the conclusions drawn from activation patching depend on both the intervention and the evaluation metric (Hase et al., 2023; Zhang & Nanda, 2024; Tan et al., 2024). Causal-abstraction approaches provide a complementary framework for testing mechanistic hypotheses through interventions (Geiger et al., 2021; Chan et al., 2022). These results make intervention alignment part of the causal claim itself. Changing an internal representation or a local score is not equivalent to changing the computation that determines the model's eventual output.

Even when an intervention reaches a behaviorally relevant computation, its effect need not persist. The Hydra Effect shows downstream computation changing in response to an earlier ablation, and subsequent work documents self-repair across model families (McGrath et al., 2023; Rushing & Nanda, 2024). For an acting model that continues generating, the analogous issue is temporal. A perturbation may change the action selected at one decision point, while a later decision recomputes the choice and restores the original action. Immediate causal control and control of the final committed behavior are therefore distinct empirical claims.

These literatures ask what information is represented, how competing evidence is resolved, how current state is maintained, and whether an identified representation causally controls behavior. We focus on the evidence required to move between these claims, treating decodability, authoritative currentness, behavioral use, and causal control as distinct measurements in agents whose task state changes during execution.

3 What Counts as Agent State?

We study task state in an acting agent whose environment changes over the course of a trajectory. Information that was correct earlier may remain available after the task has changed, and retaining such information is not itself an error. A useful account of current state must therefore separate three questions. What information is available, which value reflects the current task, and which value guides the agent's behavior? In ordinary trajectories the current authoritative value is often also the most recently presented value, so a readout may appear to track the current state while simply following the latest relevant evidence. We use D, C, and U as separable measurements of task-state tracking (Figure 1).

Decodable information (D). A task-state variable Z_t is decodable when its value can be reliably recovered from the model's hidden activations. Decodability establishes that information about the variable is accessible in the representation. It does not establish which value the model treats as current. If current and stale information agree, a successful readout is compatible with both interpretations.

Authoritative currentness (C). Currentness concerns whether a readout continues to identify the value that is currently authoritative when more recent evidence says otherwise. Let AUTH denote the authoritative current value and LAST the most recently presented value relevant to the same task variable. When AUTH $=$ LAST, following authority and following recency are observationally indistinguishable. When AUTH $\neq$ LAST, the two make different predictions and can be tested separately.

We therefore learn the readout in the ordinary setting where authority and recency agree, then apply the same readout when they conflict. The test asks whether it continues to track AUTH once recency provides a competing signal. Failure of this test does not show that information about AUTH has disappeared. The authoritative value may remain accessible elsewhere in the representation. Refitting a new readout within each condition would instead test whether AUTH remains decodable in that context. C asks whether the readout that appeared to track the current state remains valid when authority and recency diverge.

Behavioral use (U). A separate question is whether the agent's behavior follows the authoritative value. When current and stale information imply different actions, the agent's natural action reveals which value is reflected in behavior. Currentness and behavioral use can therefore differ. A readout may fail the currentness test even when the agent's action continues to follow the authoritative value, or the reverse may occur.

The three measurements support different empirical claims and do not form an ordered hierarchy. In particular,

$$D \not\Rightarrow C, \qquad C \not\Rightarrow U, \qquad U \not\Rightarrow C.$$

The decomposition makes information availability, authoritative currentness, and behavioral use separately testable rather than treating them as interchangeable evidence of agent state.

The $D/C/U$ measurements are observational. Causal interventions address a further question of whether changing hidden activations changes behavior, distinguishing an immediate change during generation from one that persists to the final action. Section 4 describes how these measurements and interventions are implemented experimentally.

4 Experimental Design

Experimental setting. We run language models as live coding agents on software-engineering tasks in real repositories, interacting with the repository over multiple turns. Our primary experiments use Qwen3.6-35B-A3B. The broader study covers six models and seven task-state families, giving 42 model–family combinations (Table 1; model details in Appendix A).

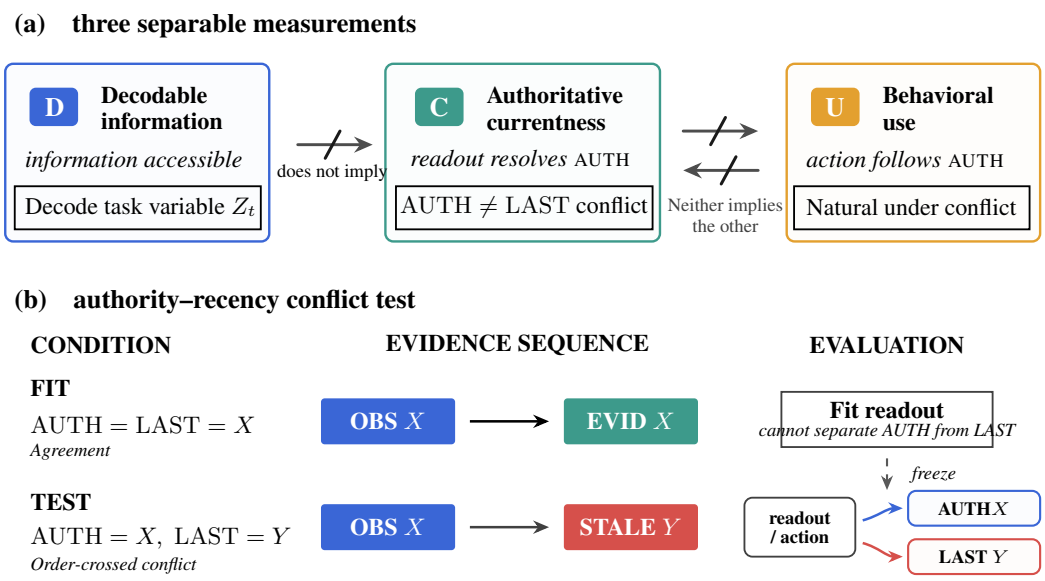

Figure 1: **Separating information, currentness, and behavioral use.** (a) D asks whether task information is decodable, C asks whether an agreement-fitted readout stays bound to the authoritative value under conflict, and U asks whether natural behavior follows that value. (b) Agreement cases (AUTH $=$ LAST) fit the readout but cannot distinguish authority from recency, while order-crossed conflict (AUTH $\neq$ LAST) makes the two predict different values.

Activations are recorded from the same forward passes that produce the agent's behavior. The repository cohorts are drawn from public open-source Python repositories on GitHub, with per-cohort repository lists in the artifact manifest. The primary file-content experiments use repository-disjoint cohorts, with 58 repositories in the decodability assay and 47 in the authority–recency conflict assay. Remaining sample sizes appear with the corresponding constructions in Appendix A.

Task-state variables. For each task-state variable, the current value is determined directly from the environment rather than from human or model annotation. Our primary setting is *file-content repair state*, where a task-relevant part of the repository is in one of two states, X or Y. The same design applies to the six further families listed in Table 1. All state contrasts are presented through observable model input or tool output. Appendix A describes each construction.

Authority–recency conflict. The decodability assay uses naturally occurring trajectory turns, and the conflict assays are controlled episodes grounded in real repositories and rendered in the same coding-agent harness. We use matched cases that differ only in the order of the current observation and stale evidence. We denote the current value established by the authoritative observation as W, and the value asserted by the stale evidence as S. Agreement cases have $W = S$, whereas conflict cases have $W \neq S$. Presentation order then determines whether AUTH and LAST coincide. In the agreement condition, stale evidence appears first and the current observation later, so AUTH $=$ LAST, and these cases fit the linear readout. In the conflict condition, the current observation appears first and stale evidence asserting the alternative value appears later, so AUTH $\neq$ LAST. The surrounding context and measurement positions are held fixed across the two conditions, so the conflict condition separates a readout that follows the current environment from one that follows the most recently presented value. In a further construction, authority must be inferred from provenance rather than explicit labels. We independently vary W, S, and presentation order, with no wording that labels either source as authoritative, current, stale, or superseded, and with repository-disjoint fitting preserved. This design tests whether the recovered readout tracks the current world rather than a fixed evidence position.

Measuring D, C, and U. We measure decodable information with linear readouts evaluated on held-out repositories. The primary experiment also includes a control based only on trajectory progress, and the broader study uses the same predefined decodability rule across cells. We mea-

sure authoritative currentness by fitting the readout on agreement cases and applying it, without refitting, to conflict cases, testing whether its predictions continue to follow AUTH rather than LAST (Section 3). We measure behavioral use from the agent's natural action on matched conflict cases, scored against the action implied by AUTH. Cases in which both values imply the same action are excluded, and explicit state reports are a secondary diagnostic only. Appendix B gives the full scoring rules and statistical criteria.

Causal interventions. The causal experiments distinguish changes in an internal decision score, changes in the action produced during generation, and changes in the final action, measured separately. In the requirement-satisfaction experiment, the same calibrated intervention is applied both before the model is choosing between the relevant answer tokens and at the forward pass in which those alternatives are scored as possible outputs, with norm-matched orthogonal controls and a held-out behavioral replication. In the file-content repair experiments, hidden activations are replaced with activations from the opposite task condition while the model is producing the relevant action, and we follow any later decisions between the same actions to determine whether the change persists or the model returns to its original choice. Intervention positions, controls, sample-selection rules, and additional validation checks are reported in Appendix E.

Evaluation protocol. Probe fitting and evaluation use disjoint repositories, and uncertainty is estimated at the repository level. Measurement positions and primary scoring rules are fixed before the corresponding cross-model and cross-family results are examined. Additional construction details, robustness analyses, and reproducibility information appear in Appendices A–E.

5 Identifying Current State Under Conflict

5.1 Decodability and Currentness

We begin with the primary file-content repair experiment. The task-state variable is strongly linearly decodable. The best repository-disjoint probe reaches balanced accuracy 0.970, compared with 0.563 for the progression-only control (Figure 2a), showing that task information is readily accessible in the model's hidden activations.

The currentness result is different. In the agreement condition, where AUTH $=$ LAST, the readout's performance peaks at balanced accuracy 0.716 across the fixed layer set. Once authority and recency are separated, the same readout falls to approximately chance with respect to AUTH at the frozen decision threshold at every tested layer. Its ranking of the authoritative value nevertheless remains substantial, reaching AUROC 0.861 at L31, so the primary failure is calibration-limited rather than a loss of AUTH ranking (Appendix C.4). Restoring the authoritative observation as the most recent evidence restores the original performance, ruling out a generic effect of the modified context. Strong decodability therefore does not establish authoritative currentness. Because authority and recency coincide during fitting, agreement data alone cannot determine which one the readout tracks. The same cell is positive under the calibration-tolerant secondary definitions in Appendix D.4.

5.2 Currentness and Behavioral Use

The authoritative value remains linearly accessible within conflict. A readout fitted directly in the conflict condition recovers AUTH with balanced accuracy 0.896, and a cross-validated intercept-only refit of the transported readout restores 0.740. A geometric analysis explains why the transported readout fails. The part of the readout that ranks AUTH mixes W and S rather than reflecting W alone (Appendix C.4).

Separating W and S during fitting recovers a readout of the current value. A readout trained to predict W on the crossed design is evaluated on held-out conflict data without changing its direction, polarity, or threshold. In the primary cell, balanced accuracy rises from the agreement-trained readout's 0.500/0.542 to 0.917/0.948 under identical folds (Appendix C.5). The recovery extends to every eligible primary C^- cell, including ranking-level failures rather than only calibration-limited ones. M2-F1 reaches crossed-fit balanced accuracy 0.842/0.789, and M1-F7 rises from AUROC 0.000 at both endpoints to 1.000 (Appendix C.6).

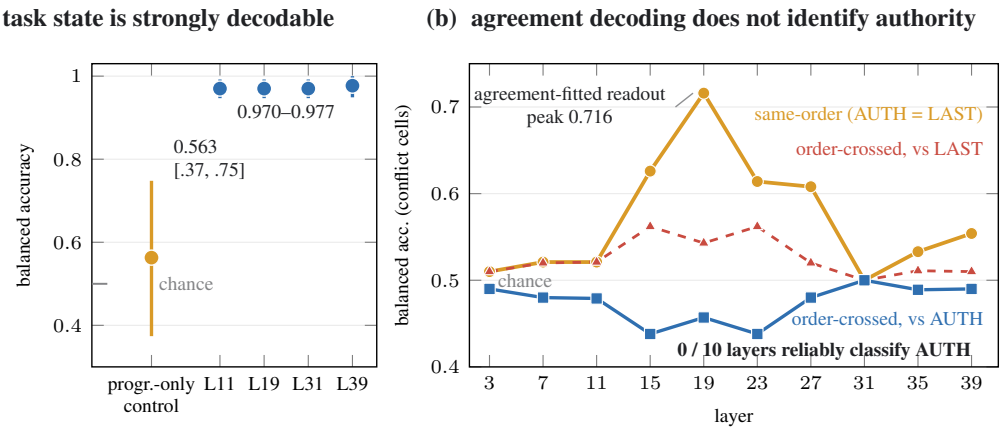

Figure 2: **Decodability does not establish authoritative currentness.** (a) Repository-disjoint probes strongly decode file-content state, while the progression-only control remains near chance; within the conflict capture, W reaches AUROC 0.972 at L31 (Appendix C.4). (b) An agreement-trained readout reaches balanced accuracy 0.716 when AUTH = LAST but fails to classify AUTH reliably under conflict. A W readout fitted with W and S varied independently transfers to held-out conflict (BA 0.917/0.948, Section 5.2).

The recovery is not explained by a fixed evidence position. In a factorial that independently varies W, S, and presentation order, authority must be inferred from provenance rather than explicit labels. A W readout fitted across both orders remains strong on held-out repositories when current evidence appears first and when stale evidence appears first. At L31, AUROC is 0.968 and 0.945 in the two orders, with balanced accuracy 0.891 in both (Appendix C.7). The construction contains no wording that labels either source as authoritative, current, stale, or superseded. Thus the jointly fitted readout identifies the current value across presentation orders rather than simply reading a fixed block position.

A text baseline that represents the two evidence blocks separately also benefits from crossed fitting. Its conflict AUTH performance reaches AUROC 0.896 and balanced accuracy 0.792. The latent readout remains significantly stronger on the same rows, with AUROC 0.975 and balanced accuracy 0.917. The paired repository bootstrap gives a latent advantage of $+0.081$ AUROC $[0.034, 0.132]$ and $+0.128$ balanced accuracy $[0.053, 0.202]$. Surface text that keeps the two blocks separate therefore captures part, but not all, of the signal the latent readout recovers.

Behavior dissociates further from currentness. On matched conflict cases, the agent's natural action follows AUTH with accuracy 0.989 $[0.967, 1.000]$. The primary experiment therefore has the regime D^+, C^-, U^+. Decodability, currentness, and behavioral use provide different evidence about agent state.

5.3 Cross-Model and Cross-Family Regimes

We next apply the same measurements to 42 combinations of six models and seven task-state families (Table 1). All 42 are D^+ under the breadth decodability criterion, so the currentness and behavioral-use comparison is interpretable in every cell. Under the frozen rule, the matrix contains 28 C^+U^+, 7 C^-U^+, 3 C^+U^-, and 4 C^-U^- cases. Exact regime counts depend on the operational currentness rule (Appendix D.4). The matrix is used to establish the existence of the four combinations rather than their prevalence.

Readout failures differ in form. In some cells the transported ranking survives while only the calibration fails, and in others the ranking itself fails or inverts (Appendices D.4 and D.5). M2-F1 provides a ranking-level transport failure, showing that the phenomenon is not confined to calibration. Requirement satisfaction is the only family that remains C^+U^+ across all six models.

Across the matrix, the consistent result is the dissociation among the measurements. Decodability does not determine whether a readout continues to track the authoritative value under conflict, and currentness does not determine whether the agent's natural behavior follows AUTH.

	state family	M1	M2	M3	M4	M5	M6
F1	file-content repair	C^-U^+	C^-U^+	C^-U^-	C^+U^+	C^+U^+	C^+U^+
F2	requirement satisfaction	C^+U^+	C^+U^+	C^+U^+	C^+U^+	C^+U^+	C^+U^+
F3	active execution binding	C^-U^+	C^+U^+	C^+U^+	C^+U^+	C^-U^+	C^+U^+
F4	execution outcome	C^+U^+	C^+U^+	C^+U^+	C^-U^-	C^+U^-	C^+U^+
F5	runtime resolution	C^-U^+	C^+U^+	C^+U^+	C^+U^+	C^-U^+	C^+U^+
F6	patch applicability	C^+U^+	C^+U^+	C^-U^-	C^+U^+	C^+U^+	C^+U^+
F7	executed-code provenance	C^-U^+	C^+U^+	C^+U^+	C^-U^-	C^+U^-	C^+U^-

Table 1: Cross-model and cross-family regimes under the primary currentness rule (model identities in Appendix A.1). Mechanically grounded state is decodable (D^+) in all 42 measurable cells, and all four C/U combinations remain present under alternative operationalizations (Appendix D.4). The examined U^- cells use constant actions rather than systematically following LAST, and exact regime counts depend on the operational C rule. (Primary F1 uses the per-layer assay; the breadth table uses aggregate held-out decoding; both give D^+.)

6 Causal Control and Persistence

The causal experiments address a question separate from the observational $D/C/U$ measurements, namely whether changes in hidden activations affect the agent's behavior. An intervention must first reach the computation that produces the action, and any resulting change must persist if it is to determine the final action.

6.1 Local Action Control

The primary file-content repair experiment shows that the action produced during generation can be changed by intervening on the model's hidden activations (Table 2). In M1–F1, opposite-condition replacement at the decision point produces the opposite-action token more often than the same-condition control (13/30 versus 5/30), although the full-vector transplant changes some token frequently in both arms (24/30 versus 20/30). None of these changes persists to the final action. The requirement-satisfaction experiment (M5-F2) shows why intervention position matters. At the original scoring position, a calibrated intervention reverses the logit difference between the two candidate answers in most cases without changing any final answer. At that point, the two answers are not yet competing outputs. Applied instead on the forward pass that produces the answer, the same intervention changes the final answer, and the effect replicates in a pre-specified held-out sample while norm-matched orthogonal controls produce no reversals. A runtime-resolution control (M5-F5) does not show the same effect, indicating that this relationship does not extend automatically across task-state families. Additional controls are reported in Appendix E.

A separate factorial experiment disentangles state from answer-token identity. At L31 the state and answer directions are nearly orthogonal (cosine -0.045), with own-factor AUROCs 0.879 and 0.987. At the live answer-producing pass, we move the activation along the state direction by the natural difference between the two state classes. This shifts the answer margin by a median 6.38 logits toward the opposite-state answer on 75/85 rows. The answer direction and norm-matched orthogonal controls remain near zero. No arm crosses the answer-margin boundary or changes the emitted answer at this scale (Appendix E.7). The result separates decodability, causal influence on the decision process, and behavioral reversal.

6.2 Decision Persistence and Self-Repair

Changing an action during generation does not guarantee that the change survives the trajectory. In M1–F1, repeating the intervention across four successive decisions raises the rate of opposite actions, but the same-condition control also produces opposite actions under clamping and the final action barely moves in either arm (Table 2), so the final-action effect is not clearly beyond the disruption floor.

The traces show M1 revisiting the same decision and restoring its original action within 1–29 generated tokens after intervention stops. In M4, where renewed decisions are rarer, roughly half of

cell	intervention site	emitted action changed	final action changed	matched control
M1-F1	action-producing pass	13/30 target-directed (24/30 raw)	0/30	SAME 5/30 (raw 20/30), paired Δ $+0.27$ $[0.09, 0.46]$
M1-F1	clamped, 4 decisions	19/26	1/26	SAME emitted 13/26 / final 0/26, paired Δ $+0.23$ $[0.00, 0.43]$
M4-F1	action position	18/18	8–10/18	SAME final 0/18, paired Δ $+0.556$ $[0.389, 0.737]$
M2-F1	action position	16/40 target-directed	16/16 of those	SAME emitted 6/40, paired Δ $+0.25$ $[0.07, 0.43]$
M5-F2	prompt-final	0/26 (margin crossed 16/26)	0/26	ORTH 0/78
M5-F2	deciding pass	13/26	13/26	ORTH 0/1080
M5-F5	deciding pass	0/26 (margin $\approx$ floor)	0/26	ORTH 0/520

Table 2: Causal evidence across cells. Interventions at the action-producing pass can change the emitted action; later decisions can restore the original choice. For M1-F1, raw token changes (24/30 vs. 20/30) show disruption in both arms; the target-directed subset carries the condition-specific effect. Paired Δ values are TARGET-minus-SAME differences with 95% repository-clustered bootstrap intervals; the clamped final-action Δ is $+0.04$ $[0.00, 0.13]$. Full controls are in Appendix E.

the changes persist; in M2, no later choice occurs. Later decision opportunities can therefore permit restoration of the original action.

7 Discussion

Our results argue against treating agent state as a single property established by one successful measurement. When authority and recency coincide during fitting, the fitted readout need not distinguish them; separating current from stale evidence recovers a readout that transfers under conflict, including when presentation order varies and explicit authority labels are removed, and a text baseline recovers some, but not all, of this information. The $D/C/U$ decomposition therefore distinguishes information availability, currentness, and behavioral use rather than letting decodability stand in for all three.

When current and stale information are correlated in ordinary trajectories, a readout trained only on those trajectories may not identify which internal signal corresponds to the current world, even when such a readout exists. This matters for agents that maintain memories or summaries, which must determine not only what information is relevant but which earlier information has been superseded.

The causal results add a further separation. Decodability need not predict causal relevance, and influence on the decision process need not be strong enough to change the chosen action; a changed action may still be reversed by a later decision.

Our experiments focus on mechanically verifiable task states; future work should extend the framework to uncertain or internally constructed state and other long-running agents. A probe intended to track current state should be trained on data that separate current from superseded evidence and tested when the two disagree.

8 Conclusion

We studied when decodable task information provides evidence of agent state in live coding agents. When authority and recency diverge, decodability, currentness, and behavioral use can dissociate. Separating current from stale evidence during fitting recovers a readout that transfers across presentation orders, including when authority must be inferred from provenance. Causal interventions further show that changing an internal quantity, the action being generated, and the final action are distinct outcomes.

AI USE STATEMENT

Generative AI tools were used as interactive research assistants throughout the project. During brainstorming, the authors engaged in extensive back-and-forth discussion with AI tools to clarify ideas, identify assumptions, organize competing explanations, and sharpen the framing of the research questions. AI tools were also used to survey related work and help assess the novelty and positioning of the proposed contributions, with the relevant literature subsequently checked by the authors. During experimentation, they assisted in refining experimental plans, suggesting controls and diagnostic analyses, and supporting implementation and debugging, while experimental choices and interpretations were made and reviewed by the authors. During manuscript preparation, AI tools helped organize the presentation, draft and revise passages, and improve clarity and concision. All reported results, citations, interpretations, claims, and final manuscript content were reviewed and validated by the authors, who take full responsibility for the work.

REFERENCES

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1), 2022.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldwosky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: A method for rigorously testing interpretability hypotheses. *Alignment Forum*, 2022.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. STALE: Can LLM agents know when their memories are no longer valid? *arXiv preprint arXiv:2605.06527*, 2026.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Xinyi Fan, Miri Liu, Ruozhen Yang, Siru Ouyang, and Jiawei Han. Can agent memory systems track evolving state? *arXiv preprint arXiv:2608.19652*, 2026.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring latent world states in language models with propositional probes. In *The Thirteenth International Conference on Learning Representations*, 2025.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, 2024.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, 2023.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. Implicit representations of meaning in neural language models. In *Proceedings of ACL*, 2021.

Belinda Z. Li, Zifan Guo, and Jacob Andreas. (How) Do Language Models Track State? In *International Conference on Machine Learning*, 2025.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023.

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*, 2024.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations, 2023.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of UIST*, 2023.

Vedant Patel. Supersede: Diagnosing and training the memory-update gap in LLM agents. *arXiv preprint arXiv:2606.27472*, 2026.

Nikolai Rozanov and Marek Rei. StateAct: Enhancing LLM base agents via self-prompting and state-tracking. In *Proceedings of the First Workshop for Research on Agent Language Models (REALM 2025)*, 2025.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42836–42855. PMLR, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

André Silva, Han Tu, and Martin Monperrus. Latent programming horizons in coding agents. *arXiv preprint arXiv:2607.05188*, 2026.

Smitha Muthya Sudheendra and Jaideep Srivastava. When decodability is not enough: Logical validity representations, behavioral dissociation, and causal tests in language models. *arXiv preprint arXiv:2609.02438*, 2026.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, 2024.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024.

A Experimental Setup and Assay Construction

A.1 Acting models and online activation capture

Six models are measured. The primary model (M1) is Qwen3.6-35B-A3B, a mixture-of-experts reasoning model that emits an internal thinking span delimited by `</think>` before its answer, run as a software-engineering agent over real repository tasks. The remaining five are Qwen3-Coder-Next (M2), DeepSeek-Coder-V2-Lite-Instruct (M3), Phi-4 (M4), GLM-4-32B (M5), and Devstral (M6). M3–M6 come from distinct model families, so the breadth study is not confined to Qwen-based architectures.

Trajectories are the agent's own multi-turn rollouts — reading files, editing, running tests, and reasoning over tool output. No teacher forcing or post-hoc re-encoding is used. Every analyzed activation was produced by the same forward passes that produced the agent's behavior. Activations are recorded at a fixed set of full-attention layers spanning the depth of each model, chosen by rule before the currentness experiments. M1 uses ten layers, L3–L39 in steps of four, and M2–M6 use per-model sets chosen under the same rule. At each measurement endpoint the per-token hidden states of the relevant span are mean-pooled to one vector per layer. Capture runs inside the serving process during generation, recording layer index, endpoint tag, and token-span metadata.

A.2 Mechanical state grounding and measurement endpoints

Task-state labels are grounded mechanically, not by annotation. The relevant region of the repository worktree (or the corresponding mechanical check for the other families, Appendix D) is evaluated per turn, so the current value of the state is a recorded property of the environment at capture time. Episodes predating the per-turn instrumentation are excluded from analyses that require per-turn labels, and no heuristic relabeling is applied.

The currentness assay reads activations at two passive endpoints, both computed by rule from the rendered prompt rather than selected by inspection. The *pre-demand* endpoint follows the last state-bearing observation and a fixed neutral continuation, before the state-dependent continuation begins. The second endpoint is the *first state-dependent generated position*. Additional endpoints were recorded during capture but the classification rules of Appendix B.4 use these two.

Development used a 50-repository split (seed 20260821), and probes requiring repository-disjoint calibration used a separate 60-repository slice with zero overlap.

A.3 Authority–recency cell construction

Each episode contributes cells crossing the stale-evidence value $S \in \{X, Y\}$ with the authoritative observation of the current world $W \in \{X, Y\}$. In the primary currentness assay, agreement cells ($S = W$) fit the decoders and conflict cells ($S \neq W$) are reserved for evaluation. The design is symmetric — each stale value appears with both observations — so stale identity and current world are orthogonal by construction. Crossing presentation order $\in \{\text{stale-first}, \text{observation-first}\}$ adds the identifying condition. In the observation-first order the stale block is the *latest* state-bearing evidence, so the authoritative value (AUTH) and the last-mentioned value (LAST) diverge on conflict cells, and a current-state readout and a recency readout make opposite predictions.

The system prompt, task scaffold, standardized literal read command on the same file region, message structure, fixed neutral continuation, endpoint definitions, and layer-capture recipe are held constant across all cells. Only the truthful state-bearing content and its presentation order differ. The primary family supplies 51 captured episodes, of which 48 are usable across 47 repositories (3 dropped for context-limit overrun, 0 reconstruction mismatches), giving 96 agreement rows per order for fitting (192 pooled) and $96 + 96$ conflict rows for evaluation, scored separately by order. The stale-first conflict cells replicate the same-order result almost exactly (peak layer 0.718 vs. 0.716, Appendix C), establishing that the same-order measurement is stable and the order-crossed reversal is attributable to the order manipulation rather than measurement drift.

A.4 OBSERVABILITY CONTROL AND CELL VALIDITY

An environment-only manipulation — differing filesystem state with byte-identical prompt text — is unobservable in principle to a text-conditioned actor, because identical causal tokens imply identical activations. This was verified mechanically (byte-identical endpoint activations in 20/20 sampled episodes), so every valid state contrast in this paper is presented through observable input.

Worked example (F1). A rendered F1 cell consists of a fixed system prompt ("You are inspecting a repository. Answer concisely and do not modify anything."), a task line ("Requirement under test: [requirement]"), and two state-bearing blocks whose order is the experimental manipulation. The authoritative block is standardized literal-read output:

```
You run the standardized read command:
$ sed -n '1,200p' [state_file]
Command output (authoritative, reflects the CURRENT contents
of [state_file]):
```python [...file contents, value X or Y...]  ```
```

The stale block is a superseded snapshot of the same region:

```
Earlier snapshot of [state_file] that you were shown
previously:
```python [...file contents, the other value...]  ```
```

In the *same-order* rendering the stale block appears first and the authoritative block last (AUTH $=$ LAST in reading order). In the *order-crossed* rendering the authoritative block appears first and the stale block last (AUTH $\neq$ LAST). The fixed neutral continuation follows the second block. The *pre-demand* endpoint is the final position of that continuation, and the *first state-dependent generated position* is the first token generated after the state-dependent demand (for the behavioral arm, the demand ends "If the requirement is currently SATISFIED by the file as it now stands, reply exactly RUN_TESTS. If it is currently VIOLATED and needs an edit, reply exactly EDIT_FILE."). Repository-specific content is elided with [. . .], and all other text is verbatim from the frozen harness.

No conflict cell is scored unless the authoritative and last-mentioned values are verified to diverge in the rendered prompt. This held on $96/96$ (latent assay) and $192/192$ (behavioral assay) conflict cells of the primary family. Episodes whose largest cell exceeded the context budget were excluded before capture by pre-tokenizing that cell (margin 512 tokens), so no included cell was truncated.

A.5 HARDWARE AND RUNTIME

Hardware is fixed per experimental series. The primary-model measurement experiments ran on H200 MIG 3g.71gb $\times 2$ with tensor parallelism 2, while the cross-family and cross-model profiles, and the causal-validation experiments on GLM-4-32B, ran on B300-class accelerators. Activations from different accelerators are not bitwise comparable and are never mixed into any contrast. Every fitted decoder is applied only to activations from its own series' hardware. All capture runs use the same serving stack version across a series, and serialized-decoder replay checks are bit-identical before any conflict evaluation (Appendix B).

B READOUT AND EVALUATION PROTOCOL

B.1 DETERMINISTIC LINEAR READOUTS

The primary currentness readouts are linear logistic-regression readouts on mean-pooled per-layer activations. They are fitted on agreement cells and evaluated on conflict cells under repository-grouped cross-validation. Diagnostic readouts fitted under other constructions are identified explicitly below. Model-specific decoders are fitted on the same calibration splits. There is no per-family or per-model tuning of layers, pooling, endpoints, or probes.

Decoder fitting is deterministic, using single-threaded execution, float64 arithmetic, a solver run to convergence, and serialization of the decoder at fit time. Downstream analyses load the serialized

decoder rather than refitting, and serialization-based replay reproduces every statistic bit-identically. Because nondeterministic fitting can move individual-layer statistics by up to approximately 0.04 balanced accuracy, conclusions are based on layer-set-level patterns rather than marginal layer identities, which are reported descriptively only.

B.2 Repository-disjoint splits and statistical conventions

Probe fitting and evaluation are repository-disjoint throughout. Primary uncertainty is the repository-clustered bootstrap confidence interval. Where a pre-specified analysis also specifies a within-case two-sided sign test, disagreement between the two statistics is resolved to the weaker interpretation. Denominators are fixed before scoring, and mechanically unscoreable cells are preserved as missing, never imputed or substituted. A positive behavioral result on any probe requires the clustered CI lower bound to exceed 0.5, not merely a higher point estimate.

B.3 Behavioral scoring and the parse gate

The behavioral assay uses the corresponding order-crossed conflict construction and matched row identities in a separate text-only run, with no activation capture. Two probes are scored mechanically against both AUTH and LAST, the naturally state-dependent next action (primary) and an isolated explicit current-state query (secondary). Query replies are matched against the two candidate values after normalization. Replies naming neither value, or hedging between both, score as incorrect against both baselines. Action correctness is mechanical. The emitted action is correct iff it is the action that the authoritative value calls for, and cells without a mechanically decidable correct action are excluded from the denominator.

Because the primary acting model emits a reasoning preamble before its answer, replies are parsed from the post-`</think>` span with a generation budget large enough to clear the preamble, and a hard precondition requires $\geq 90\%$ of replies to parse before any scoring executes. An assay that fails the gate is reported as unmeasured rather than scored on absent measurements. The canonical behavioral run passed the gate at $177/192 = 0.922$. The same parse gate applies per cell in the cross-model profiles. A cell whose parse rate fails the gate is reported as unmeasured for the affected property rather than scored. A constant (degenerate) action policy is treated differently. It is a measured behavioral outcome, not a scoring failure, and receives U^- under the primary rule of Appendix B.4, with the robustness recount of Appendix D.4 treating such cells as indeterminate instead.

B.4 $D/C/U$ classification rules

Each (family, model) cell receives three boolean verdicts under rules fixed before the cross-model measurements:

- D **(decodable information).** Positive when held-out agreement-cell decoding clears the frozen decodability criterion (repository-clustered CI lower bound above chance). In the primary F1 assay, we additionally compare against the progression-only nuisance control. D functions as an availability control. The state families differ in how close the mechanical label sits to surface evidence, so a D^+ verdict certifies that the information is present, not that decoding it is difficult.
- C **(authoritative currentness).** The agreement-fitted readout is transported without refitting to order-crossed conflict cells and scored against AUTH at the two passive endpoints of Appendix A.2. Under the primary (disjunction) rule, a cell is C^+ if the transported readout aligns with AUTH (clustered CI lower bound > 0.5) at either endpoint, and C^- if it aligns at neither. Alternative operationalizations are reported in Appendix D.4; the existence of all four C/U regimes is invariant to the choice of rule, and exact regime counts are not.
- U **(behavioral use).** Positive iff the natural state-dependent action follows AUTH on the matched behavioral conflict instances with clustered CI lower bound > 0.5 and the parse gate passes. Cells whose action policy is constant across conflict values are U-degenerate and receive U^- under the primary rule. The degenerate-excluded recount of Appendix D.4 treats them as indeterminate instead.

These rules classify measurements. They do not assert mechanisms. In particular C^- is a statement that the agreement-trained readout no longer provides reliable current-value predictions under conflict, not a statement about the absence of authoritative-state information (Appendix C.3), and the diagnostically distinct readout-failure modes behind C^- are profiled separately in Appendix D.5.

B.5 Calibration-tolerant currentness diagnostics

Two secondary currentness rules separate directional transport from calibration stability. Both were fixed before their outcomes were inspected, and both preserve the primary rule's endpoint-combination convention (positive at either passive endpoint).

AUROC-C. The transported agreement-trained direction and its polarity remain frozen, and the readout's ranking of AUTH is scored on conflict rows with repository-clustered uncertainty. A cell is positive iff the clustered CI lower bound on the AUROC exceeds 0.5 at either endpoint. No $\max(\text{AUROC}, 1 - \text{AUROC})$ or polarity rescue is applied. A direction that ranks the authoritative value below chance scores below chance.

Intercept-C. The direction and polarity remain frozen. Only a scalar intercept is recalibrated, on conflict rows from training repositories, and evaluation uses held-out repositories only, with out-of-fold predictions pooled for the balanced-accuracy estimate and its clustered CI. A cell is positive iff the pooled held-out CI lower bound exceeds 0.5 and the balanced accuracy reaches 0.60 at either endpoint. The 0.60 floor is the pre-specified effect-size floor of the recount protocol, fixed before any outcome was inspected.

Together the two diagnostics distinguish a readout whose fitted direction still ranks the authoritative value under conflict but whose decision boundary fails (*calibration-limited* transport failure) from one whose ranking, with direction and polarity frozen, itself fails (*ranking-level* transport failure). They are secondary operationalizations, and the pre-specified primary C rule of Appendix B.4 remains unchanged. Results appear in Appendix D.4.

C Primary-State Measurement Results

All results in this appendix concern the primary cell, file-content repair state (F1) on the primary acting model (M1).

C.1 Full per-layer decodability record (D)

The decodability assay uses matched turns with 45 balanced pairs over 58 repositories under repository-disjoint 5-fold cross-validation.

probe	balanced accuracy	95% CI
progression-only (4 vars, no latents)	0.563	[0.374, 0.748]
latent L11	0.970	[0.947, 0.992]
latent L19	0.970	[0.947, 0.992]
latent L31	0.970	[0.947, 0.992]
latent L39	0.977	[0.949, 1.000]

A wiring check confirmed the per-layer vectors are distinct (four distinct norms, zero byte-identical pairs, pairwise cosine 0.48–0.93).

Shallow lexical baseline. A frozen TF-IDF logistic-regression baseline, fitted on the rendered prompt text under the same repository-disjoint splits, reaches balanced accuracy 0.813 (AUROC 0.889 $[0.83, 0.94]$) on agreement cells, against 1.000 for the latent readout. Transported to order-crossed conflict cells, the lexical readout scores exactly 0.500 against AUTH, while the latent readout retains substantial authoritative-value ranking (Appendix C.4). The state is therefore partly surface-accessible, which is consistent with D's role as an availability measurement, but latent decoding is stronger, and the tested lexical readout does not transport to authoritative conflict.

C.2 SAME-ORDER AND ORDER-CROSSED CURRENTNESS SWEEPS (C)

The same-order condition is fitted on agreement cells and evaluated on conflict cells, with the world label given by the authoritative observation. Balanced accuracy with 95% CIs follows.

layer	pre-demand	first state-dependent position
L3	0.510 [0.500, 0.530]	0.501 [0.470, 0.533]
L7	0.521 [0.500, 0.543]	0.480 [0.440, 0.500]
L11	0.521 [0.500, 0.543]	0.530 [0.500, 0.570]
L15	0.626 [0.582, 0.670]	0.542 [0.511, 0.573]
L19	**0.716** [0.613, 0.814]	0.571 [0.530, 0.620]
L23	0.614 [0.571, 0.657]	0.562 [0.530, 0.597]
L27	0.608 [0.550, 0.699]	0.592 [0.540, 0.650]
L31	0.500 [0.500, 0.500]	0.542 [0.521, 0.554]
L35	0.533 [0.500, 0.578]	0.541 [0.510, 0.572]
L39	0.554 [0.510, 0.599]	0.542 [0.521, 0.554]

The same-order signal is carried by mid-depth layers (L15–L27). The 0.716 value is the descriptive maximum over the fixed layer sweep, and no multiplicity-adjusted significance claim is attached to it. The next table reports order-crossed conflict cells (observation first, stale evidence most recent), scored against both candidates, with CIs shown where the lower bound clears 0.5.

layer	pre vs. AUTH	pre vs. LAST	P1 vs. AUTH	P1 vs. LAST
L3	0.490	0.510	0.521	0.479
L7	0.480	0.520	0.490	0.510
L11	0.479	0.521	0.468	0.532
L15	0.438	**0.562** [0.50, 0.61]	0.490	0.510
L19	0.457	0.543	0.447	**0.553** [0.52, 0.59]
L23	0.438	**0.562** [0.55, 0.58]	0.469	0.531
L27	0.480	0.520	0.437	**0.563** [0.53, 0.60]
L31	0.500	0.500	0.511	0.489
L35	0.489	0.511	0.501	0.499
L39	0.490	0.510	0.500	0.500

No layer aligns with the authoritative value at either endpoint (0/10). The stale-first control uses $S \neq W$ conflict rows and is distinct from the $S = W$ agreement rows used to fit the decoder. These stale-first conflict cells of the same run replicate the same-order sweep almost exactly (L15 0.624 vs. 0.626; L19 0.718 vs. 0.716; L23 0.586 vs. 0.614; L31 0.542 vs. 0.500), so the reversal is attributable to the order manipulation, not to measurement drift.

C.3 CONFLICT-CONTEXT REFIT

The transported readout's collapse could reflect absence of linear authoritative-state information under conflict, or a failure of the agreement-trained readout to transfer. To distinguish these, a deterministic decoder of identical form was fitted *on conflict cells only*. The fit used repository-disjoint 3-fold cross-validation with a fresh fold seed (no reuse of agreement folds), was endpoint-balanced across both latent endpoints, and followed the same solver protocol as all other decoders. The interpretation rule was fixed before any fit. A within-conflict CI lower bound above 0.5 establishes within-conflict linear decodability, and hence reads the transported readout's collapse as readout instability. Otherwise this diagnostic is inconclusive, because failure to establish above-chance linear decoding is not evidence that the information is absent.

In the primary cell the within-conflict decoder reaches balanced accuracy $\mathbf{0.896}$ $[0.833, 0.952]$ against AUTH, while the transported agreement-fitted readout scores ≈ 0.51 on the same cells. The authoritative value therefore remains strongly linearly decodable within the conflict context. What fails is the transfer of the agreement-trained readout, which no longer provides reliable current-value predictions under conflict. Because this decoder is fitted with conflict labels, it is a diagnostic only — within-context decodability does not establish that the agreement-trained readout transfers — and does not replace the operational currentness assay. The same diagnostic applied across models and

families, with its control cells, appears in Appendix D.5. The geometry of the agreement-trained readout itself is analyzed in Appendix C.4.

C.4 GEOMETRY OF THE AGREEMENT-TRAINED READOUT

A pre-specified offline analysis characterizes the agreement-trained readout on the frozen primary capture. The capture is a crossed design in which the current value W and the stale-evidence value S vary independently (96 cases per condition), so a direction associated with each factor can be estimated separately. Both factors are strongly decodable, with cross-validated AUROC 0.972 for W and 0.958 for S at the pre-specified primary analysis layer L31 (0.984 and 0.919 at L35).

The readout mixes both factors. The agreement-trained readout aligns substantially with both factor directions, with cosine 0.448 to the W direction and 0.428 to the S direction at L31. The two factor directions are themselves correlated (cosine 0.556). The readout therefore contains substantial components associated with both the current value W and the stale value S and is not authority-specific. Because the two factor directions are correlated, their raw projection coefficients are not directly comparable, so we characterize their joint span.

The AUTH-relevant component lies in the two-factor span. The projection of the readout onto the span of the two factor directions holds about half of its norm (0.497 at L31, 0.491 at L35), yet this projection retains nearly all of the readout's ability to rank the authoritative value under conflict (AUROC 0.838 against 0.861 for the full readout at L31; 0.886 against 0.902 at L35), while the orthogonal residual carries essentially no linear authoritative-value signal (AUROC 0.481 and 0.498). The component relevant to ranking the authoritative value in this assay lies in the two-factor span, and the residual carries structure irrelevant to this contrast.

Calibration-limited transport failure. Under agreement the readout ranks the authoritative value at AUROC 1.000. Under conflict it retains AUROC 0.861 (L31) and 0.902 (L35), while balanced accuracy at the frozen decision threshold is exactly 0.500. Recalibrating the intercept alone, with the direction frozen and calibration and evaluation on disjoint repositories, restores balanced accuracy to 0.740 $[0.67, 0.81]$ (L31) and 0.771 $[0.68, 0.85]$ (L35). The primary transport failure is therefore calibration-limited, with a real but smaller reduction in ranking quality, rather than a loss of the fitted direction. This analysis explains why the primary cell receives C^- under the frozen criterion.

quantity	L31 (primary)	L35
factor decodability, W / S (CV AUROC)	0.972 / 0.958	0.984 / 0.919
$\cos(\text{readout}, d_W)$ / $\cos(\text{readout}, d_S)$	0.448 / 0.428	0.459 / 0.388
$\cos(d_W, d_S)$	0.556	0.521
projected norm fraction	0.497	0.491
conflict AUROC: full / projected / residual	0.861 / 0.838 / 0.481	0.902 / 0.886 / 0.498
conflict BA: frozen / CV intercept-refit	0.500 / 0.740	0.500 / 0.771

Presentation order is constant in this capture, so an order factor is not estimable from this analysis. The order manipulation itself is carried by the conflict assay of Appendix C.2.

C.5 READOUT OF THE CURRENT VALUE UNDER CROSSED FITTING

The geometry analysis shows that the agreement-trained readout is a mixture of the W and S factors and that its transport failure is calibration-sensitive. This raises a constructive question. When the two factors vary independently during fitting, can a readout of the current value W be identified, and does it transport? Using the same frozen crossed capture and identical repository-disjoint folds, a readout of W was fitted on the crossed design and evaluated on held-out conflict cells with its direction, polarity, and threshold all frozen at fit time. The agreement-only readout serves as the comparator under the identical folds.

	crossed-fit W readout		agreement-only readout	
	L31	L35	L31	L35
held-out W decoding (AUROC)	0.972	0.984	0.976	0.964
held-out W decoding (BA)	0.912	0.932	0.927	0.927
conflict AUTH AUROC [CI]	**0.975** [0.95, 1.0]	**0.988** [0.97, 1.0]	0.848 [0.79, 0.91]	0.898 [0.85, 0.95]
conflict frozen-threshold BA [CI]	**0.917** [0.85, 0.97]	**0.948** [0.89, 0.99]	0.500 [0.50, 0.50]	0.542 [0.51, 0.58]

When W and S vary independently during fitting, the readout of W transports to held-out conflict without any recalibration. Frozen-threshold balanced accuracy reaches 0.917–0.948 with conflict AUTH AUROC 0.975–0.988. Agreement-only fitting therefore does not reliably identify the W readout, because W and S coincide in the fitting condition, and the same folds reproduce the calibration-limited signature for the agreement-trained comparator (AUROC 0.85–0.90, frozen-threshold balanced accuracy at chance). This connects directly to Appendix C.4. The geometry shows the agreement readout is a W/S mixture whose calibration fails under conflict, and the crossed fit shows that a stable readout can be obtained from the same activations when W and S vary independently during fitting. The same readout remains calibrated on held-out agreement and conflict data. On conflict-only evaluation rows AUTH and LAST are complements, so a separate stale-value alignment column is not independently informative there, and the readout's W-specificity is established by the crossed training cells.

Matched lexical control. We repeated the crossed fitting procedure with the frozen TF-IDF logistic-regression baseline, using the same repository-disjoint folds and held-out conflict rows ($n = 96$), with no tuning of vocabulary, regularization, threshold, folds, or features. Crossed supervision does not recover transport for this shallow text baseline. Held-out W decoding reaches AUROC 0.732 / BA 0.651, but conflict AUTH remains at AUROC 0.500 and frozen-threshold BA 0.500. On the identical conflict rows, the latent crossed-fit readout reaches AUROC 0.975 and BA 0.917. Thus the constructive result is not explained by crossed supervision alone acting on the shallow rendered-text features captured by TF-IDF. A position-aware variant of the same baseline, representing the two state-bearing blocks separately under the identical crossed supervision and folds, recovers substantially more. Held-out W decoding reaches AUROC 0.894 / BA 0.792, and conflict AUTH reaches AUROC 0.896 / BA 0.792. Position-aware surface text therefore recovers substantial W information under crossed supervision, but remains below the latent readout (0.975/0.917 on the same rows). Under agreement-only fitting, the position-aware baseline also transports at chance, indicating that the authority–recency under-identification is not specific to latent representations.

The same crossed-fit test is extended to every eligible primary C^- cell in Appendix C.6.

C.6 Crossed fitting across primary C^- cells

The crossed-fit test of Appendix C.5 was applied to every primary C^- cell whose retained rows form a crossed $W \times S$ design in which both factors vary independently. Six cells are eligible — M2-F1, M5-F3, M5-F5, M4-F4, M4-F7, and M1-F7 — with M1-F1 reported by reference from Appendix C.5; M3-F6 is not included because its retained rows do not form the required crossed design, and no other primary C^- cell had retained data supporting the crossed fitting construction. For each cell, a readout trained to predict W is fitted on the crossed design under repository-disjoint folds and evaluated on held-out conflict rows without changing its direction, polarity, or threshold. The agreement-trained comparator is refitted under the identical folds and reproduces the frozen C^- signature in every cell before the crossed fit is scored.

	agreement-trained (BA / AUROC)		crossed fit (BA / AUROC)			
cell	P0	P1	P0	P1	n / repos	criterion
M2-F1	0.500 / 0.501	0.474 / 0.604	0.842 / 0.875	0.789 / 0.817	38 / 19	met
M5-F3	0.528 / 0.657	0.500 / 0.716	0.889 / 0.948	0.861 / 0.951	36 / 30	met
M5-F5	0.500 / 0.494	0.603 / 0.709	1.000 / 1.000	1.000 / 1.000	58 / 48	met
M4-F4	0.500 / 0.587	0.500 / 0.579	1.000 / 1.000	1.000 / 1.000	38 / 32	met
M4-F7	0.500 / 1.000	0.500 / 1.000	1.000 / 1.000	1.000 / 1.000	56 / 47	met
M1-F7	0.000 / 0.000	0.500 / 0.000	1.000 / 1.000	1.000 / 1.000	56 / 47	met
M1-F1	0.500 / 0.848	0.542 / 0.898	0.917 / 0.975	0.948 / 0.988	96 / 47	met

The n column gives the number of held-out decisive conflict rows evaluated per cell and the number of repositories in the corresponding capture, recovered from the frozen per-cell manifests.

The criterion, fixed before scoring, requires balanced accuracy ≥ 0.70 with clustered CI lower bound above 0.5, and every eligible cell clears it at both endpoints. The agreement comparator in this table is refitted under the same folds used for crossed fitting. The 0.374 M2-F1 AUROC in Appendix D.5 comes from the archived secondary diagnostic lineage and is therefore not directly comparable to these same-fold values. Four C^+ control cells are saturated at 1.0 under both arms, so the effect is specific to the primary C^- cells. The crossed-fit readout meets the criterion in every eligible primary C^- cell, including the three whose agreement-trained readouts lose or invert ranking. M2-F1 rises from chance to 0.842/0.789, and M1-F7 rises from AUROC exactly 0.000 at both endpoints to 1.000 on the same rows, folds, and thresholds, with only the fitting supervision changed. Because W and S are complements on these conflict rows, AUROC 0.000 against AUTH corresponds to perfect ranking of the stale value. These results show that severe conflict failures can occur in the agreement-trained readout even when a readout of W transfers reliably once W and S are separated during fitting.

C.7 Authority inferred from provenance

The primary conflict cells adjudicate authority in the rendered text. A further variant removes that adjudication entirely, so that authority must be inferred from provenance. A live tool observation supplies one source of the task-state value, conflicting ticket and developer context supplies the other, and the two sources differ only in provenance — no lexical text declares either source authoritative, current, stale, or superseded. The cohort comprises 480 rows over 48 episodes and 47 repositories, constructed and screened under a frozen template gate.

quantity	result (L31)
D (agreement, repository-CV)	BA 0.943 / AUROC 0.973
C transport vs. AUTH	BA 0.547 / AUROC 0.825
C transport vs. LAST	AUROC 0.556
natural action follows AUTH (parsed actions)	0.870 $[0.80, 0.93]$
by order (observation-first / notes-last)	0.915 / 0.816
paired Δ vs. explicit adjudication (order-matched)	-0.196 $[-0.29, -0.10]$
parse rate (vs. 0.896 matched explicit-adjudication arm)	0.563

The qualitative pattern persists without explicit authority labels. The state remains strongly decodable, and the transported readout again fails at the frozen threshold while retaining substantial authoritative-value ranking (AUROC 0.825, with no corresponding stale-value loading at 0.556). Conditional on a parseable action, AUTH-following is 0.870 (0.816 in the matched stale-last condition). Because the parse rate is only 0.563 and falls below the primary behavioral gate, these rates are descriptive and are not interpreted as unconditional behavioral use. Explicit adjudication contributes materially. On the same instances, AUTH-following is 0.196 lower without it (0.804 against 1.000 under the order-matched comparison), and the parse rate falls from 0.896 to 0.563, with the model hedging more often when no source is adjudicated.

The crossed-fit analysis of Appendix C.5 was repeated on this construction. A readout trained to predict W on this crossed design, with authority carried by provenance, transports to held-out conflict rows at AUROC 0.946 / BA 0.875 (L31; L35: 0.929/0.854), against the agreement-trained comparator's 0.825/0.547 (0.815/0.537 at L35). The crossed-fit rescue therefore replicates when authority is encoded by provenance rather than explicit adjudicative wording.

Varying presentation order. Within this cohort, a balanced subdesign crosses W, S, and presentation order independently (8 cells $\times$ 48 rows $=$ 384 rows over 47 repositories). Whether the live observation or the conflicting notes appear first is independent of which value is current, and an automated screen finds zero occurrences of authority-adjudicating wording. A W readout fitted jointly across both orders on repository-disjoint folds evaluates at AUROC 0.968 $[0.942, 0.988]$ on current-first rows and 0.945 $[0.897, 0.980]$ on stale-first rows, with balanced accuracy 0.891 in both orders. Restricted to conflict rows, AUROC is 0.958 and 0.930. The S readout is similarly order-flat (0.951/0.952). As a manipulation check, presentation order itself is perfectly decodable (AUROC

1.0), so the two orders are representationally distinct while the W readout remains order-invariant. A readout that reads a fixed block position cannot produce this pattern.

A diagnostic asks whether a readout fitted in only one order transfers to the other. Transfer is above chance on all rows (0.878 current-first to stale-first, 0.807 in the reverse direction) but partial on conflict rows (0.707 one way, 0.525 the other, against same-order references of 0.99 and 0.965). Joint-order fitting yields an order-general W readout, but a readout fitted in only one order does not necessarily transfer across order on conflict rows.

C.8 Full behavioral-use record (U)

The canonical behavioral run comprises 48 usable episodes across 47 repositories, with 192 conflict cells whose AUTH $\neq$ LAST divergence was verified on 192/192, and a parse rate of 0.922 (Appendix B.3). Each arm has 96 generations over 94 unique underlying conflict instances. Of these, $84/96$ query-arm and $93/96$ action-arm generations were scoreable, and all non-scoreable generations were empty generations, with no filtered third category.

probe	n	repos	vs. AUTH	vs. LAST	follows
explicit current-state query	84	43	0.679 [0.602, 0.756]	0.321	AUTH
state-dependent next action	93	47	$\mathbf{0.989}$ [0.967, 1.000]	0.011	AUTH

Both probes favor authority. Descriptively, authority alignment is stronger in natural action (0.989) than in explicit self-report (0.679). Because the explicit query changes the computation being measured, we do not interpret this magnitude gap mechanistically.

C.9 Decision-event transport and frame alignment

Two offline analyses on frozen representations support the interpretation of Appendix E.5. A powered transport matrix over eight recorded positions (79 rows / 50 repositories; all positions retained; local diagonals ≥ 0.944) shows that a readout fitted at any one decision event evaluates at the others near ceiling, with median $R \approx 0.99$ within the decision-event block and AUROC 1.00 from every decision event to the action-string onset. The transport losses that exist are confined to the prompt-final position rather than distributed across events. A label-free rank-≤ 32 map fitted only on paired prompt-final and first-decode states recovers 0.733 of that gap (identity 0.610 versus a same-position ceiling of 1.000), generalizes to later decision events (median 0.73; weaker at the action-string onset, 0.34), and is destroyed by pair shuffling (-0.398), consistent with a systematic frame difference between the two positions rather than a population offset. The same comparison on M2 gives a deficit of 0.027 against M1's 0.390, so the frame difference is model-specific and is not offered as an explanation of the C^- regime.

D Cross-Model and Cross-Family Results

D.1 Six models and seven families

The profile study asks whether the dissociations of Section 5 are specific to one (state family, acting model) pair. The final breadth study contains seven mechanically grounded state families, F1–F7, each profiled on the six models of Appendix A.1; each family specification and its scoring rules were frozen before its outcomes were inspected. The families are listed below, with compact ids in parentheses.

- **File-content repair state** (F1) — the primary family of Appendix C. Which of two mechanically hashed contents the task-relevant repository region currently holds.
- **Requirement-satisfaction state** (F2). Whether a named requirement is currently satisfied. The mechanical state is the pinned pytest-node result executed against the current worktree. Its authoritative block is verbatim checker output — the state is stated in near-classifier form.

- **Active execution-binding state** (F3). Which of two candidate execution targets is currently bound. The mechanical state is the execution-relevant selector value read by a fixed parser.
- **Execution-outcome state** (F4). Whether the most recent execution passed or failed. The mechanical state is the native pytest tail only, with no added classifier line.
- **Runtime-resolution state** (F5). Which dependency version is currently resolved. The mechanical state is the version string printed by the resolution command, and the actor must map pure evidence onto the pin it writes.
- **Patch-applicability state** (F6). Whether a candidate patch applies cleanly to the current worktree. The mechanical state is `git apply --check`. Its evidence contains explicit outcome-like wording, placing F6 in the same explicit-evidence regime as F2.
- **Executed-provenance state** (F7). Which of two candidate code versions was actually executed by the most recent run. The mechanical state is recorded execution provenance. Like F5, the label must be inferred from evidence rather than read off a classifier line.

Each (family, model) cell reuses the same machinery of agreement-cell fitting, order-crossed conflict evaluation against both AUTH and LAST, endpoint definitions, deterministic decoder protocol, and the behavioral assay with its parse gate (Appendices A and B). Per-family splits are repository-disjoint. A per-cell text-only viability check precedes capture, and a cell that failed it would be reported as unmeasurable and kept in the table rather than dropped. All 42 cells passed viability, and replay checks are bit-identical before any conflict evaluation. Earlier models' profiles were measured outcomes, not targets for the later models.

D.2 THE 6×7 MATRIX

D^+ holds in $42/42$ cells of the frozen matrix (held-out agreement decodability range 0.80–1.00), so cells are annotated by their C/U regime only.

model	F1	F2	F3	F4	F5	F6	F7
M1 Qwen3.6-35B-A3B	$C{-}U{+}$	$C{+}U{+}$	$C{-}U{+}$	$C{+}U{+}$	$C{-}U{+}$	$C{+}U{+}$	$C{-}U{+}$
M2 Qwen3-Coder-Next	$C{-}U{+}$	$C{+}U{+}$	$C{+}U{+}$	$C{+}U{+}$	$C{+}U{+}$	$C{+}U{+}$	$C{+}U{+}$
M3 DeepSeek-Coder-V2-L	$C{-}U{-}$	$C{+}U{+}$	$C{+}U{+}$	$C{+}U{+}$	$C{+}U{+}$	$C{-}U{-}$	$C{+}U{+}$
M4 Phi-4	$C{+}U{+}$	$C{+}U{+}$	$C{+}U{+}$	$C{-}U{-}$	$C{+}U{+}$	$C{+}U{+}$	$C{-}U{-}$
M5 GLM-4-32B	$C{+}U{+}$	$C{+}U{+}$	$C{-}U{+}$	$C{+}U{-}$	$C{-}U{+}$	$C{+}U{+}$	$C{+}U{-}$
M6 Devstral	$C{+}U{+}$	$C{+}U{+}$	$C{+}U{+}$	$C{+}U{+}$	$C{+}U{+}$	$C{+}U{+}$	$C{+}U{-}$

The census under the primary (disjunction) rule is 28 C^+U^+, 7 C^-U^+, 3 C^+U^-, and 4 C^-U^-. Several structural observations hold descriptively. Only F2 is model-invariant (stable on 6/6 models, statistics at or near 1.0 everywhere). F1 currentness is model-dependent rather than family-invariant (C^- on M1, M2, and M3, C^+ on M4, M5, and M6, and the regime label alone does not predict how a model organizes its decisions over time, as Appendix E.6 shows). F7 spans all four C/U regimes across the six models rather than exhibiting a repeatable dissociation. The C^+U^- regime (readout transports, natural action degenerates) recurs on two models, concentrated in the F4/F7 action formats. The examined U^- cells use a constant action across the conflict values rather than systematically following LAST. Cell counts are not prevalence estimates, because the families and models were selected for measurement coverage, not sampled from a population.

D.3 PER-CELL CONTINUOUS RECORDS

Boolean verdicts summarize continuous per-cell statistics, and representative full records follow. (The complete per-cell statistics for all 42 cells are recorded in the archived full-statistics artifact of the experimental series.)

M1. The held-out agreement-decodability positive controls are F2 1.000 $[1, 1]$, F3 0.771 $[0.667, 0.875]$, F4 1.000, and F5 0.961 $[0.914, 0.993]$. (F1's 0.970 is the per-layer matched-turn statistic of Appendix C.1. The breadth cells use the aggregate held-out criterion, and both classify F1 as D^+.) Conflict-cell records follow.

- **F2** (50 conflict cells / 25 repos). The passive readout tracks AUTH at 1.000 $[1, 1]$ at both endpoints (LAST at 0.000), and the action is correct in 47/47 scoreable cells ($p = 1.4 \times 10^{-14}$, 3 unparseable replies reported). No dissociation.
- **F3** (36 conflict cells / 18 repos). The passive aggregate sits at exact chance against *both* baselines (0.500 pre-demand with a degenerate CI, $[0.417, 0.583]$ at the second endpoint), while the action is correct in 36/36 cells ($p = 2.9 \times 10^{-11}$). This is a dissociation, and the conflict readout collapses to chance on both baselines rather than being recency-dominated.
- **F4** (38 conflict cells / 19 repos). Passive AUTH alignment is 0.895 $[0.789, 0.974]$ pre-demand and 0.684 $[0.579, 0.789]$ at the second endpoint, and the action is correct in 38/38 cells ($p = 7.3 \times 10^{-12}$). No dissociation — notably with a native test tail only.
- **F5** (58 conflict cells / 29 repos). Decodability is near ceiling (0.961) while the passive readout sits at exact chance on both baselines pre-demand (second endpoint 0.552 $[0.500, 0.603]$, lower bound not clearing 0.5), and the action is correct in 58/58 cells ($p = 6.9 \times 10^{-18}$). Dissociation.

M2. The table reports D, C at the pre-demand and second endpoints, and U per family.

family	D (held-out agreement)	C: passive vs. AUTH	U: action vs. AUTH
F1	0.926 $[0.878, 0.965]$, n=376	0.521 / $0.532 \rightarrow C-$	0.606 $[0.53, 0.68]$, p=0.049
F2	1.000 $[1, 1]$	1.000 / $1.000 \rightarrow C+$	1.000 $[1, 1]$, p=1.8×10^{-15}
F3	0.979 $[0.938, 1.000]$	0.667 / $\mathbf{0.972}$ $[0.92, 1.00] \rightarrow C+$	1.000 $[1, 1]$, p=2.9×10^{-11}
F4	1.000 $[1, 1]$	0.658 / 0.711 $[0.61, 0.82] \rightarrow C+$	0.711 $[0.61, 0.82]$, p=0.014
F5	1.000 $[1, 1]$	$\mathbf{0.983}$ $[0.95, 1.00]$ / $0.862 \rightarrow C+$	0.690 $[0.60, 0.78]$, p=0.0054
F6	0.993	1.000 / $0.875 \rightarrow C+$	0.875
F7	1.000	1.000 / $1.000 \rightarrow C+$	1.000

Model-2 behavioral-use margins are heterogeneous. They are significant in all families but substantially weaker than M1's in F1, F4, and F5, and descriptively M2 follows stale evidence more often than M1. The frozen M2-F1 U verdict is carried by the stale-last matrix cells (0.606 $[0.53, 0.68]$, n=94). A separate carrier factorial on a different template and cohort additionally shows order dependence in this cell. The probability that the action follows the authoritative value is 0.861 when the stale block comes first and 0.486 when it comes last. This diagnostic is not the source of the frozen U verdict. A balanced factorial test on the same cell also shows that the authoritative value stays fully decodable in M2 (balanced probe 0.95–0.97), so its C^- verdict is not an absence of authority information. F3's passive alignment sharpens between endpoints ($0.667 \rightarrow 0.972$), consistent with stronger authority alignment at the first generated position than at rest.

M3 (DeepSeek-Coder-V2-Lite-Instruct). F2–F5 and F7 are $C+U+$. Two cells are $C-U-$. In **F1**, the transported passive readout is unstable under conflict and natural action *also* fails, with the action following AUTH at exactly 0.500 ($n = 94$ decisive conflicts, sign test $p = 1.0$). **F6**'s failure is calibration-level (Appendix D.5). M3 therefore does not exhibit the passive-readout–behavior dissociation on F1, because readout and behavior fail together.

M4 (Phi-4). F1 — the family that fails on M1, M2, and M3 — is $C+U+$ here, with passive authority accuracy 0.691 $[0.628, 0.766]$ at the second endpoint and action 0.75 ($p = 5 \times 10^{-6}$). The deviating cells are **F4** and **F7**, both $C-U-$ with degenerate transported readout and natural-action measures; the F4 diagnostic record is mixed-intermediate (transported decoder predicts a single class, predicted-positive fraction 0.00, while ranking weakly above chance, AUROC 0.618; intercept-only recalibration recovers only partially, 0.684; conflict-context refit 1.000).

M5 (GLM-4-32B). F1, F2, and F6 are $C+U+$. F3 and F5 occupy the same $C-U+$ regime as the primary M1-F1 dissociation, and the M5-F5 diagnostic record is a clear ranking-level case (Appendix D.5). F4 and F7 are $C+U-$ — the inverse dissociation, in which the agreement-fitted readout transports but the natural action degenerates. F7's C^+ verdict is endpoint-resolved (transported accuracy 0.500 pre-demand, 1.000 at the first state-dependent position).

M6 (Devstral). Six families are $C+U+$. F3 is borderline-C^+ (transported accuracy $0.639/0.639$, CI lower bounds $0.53/0.50$), and its diagnostics resolve it as calibration-level with the direction intact (AUROC $0.96/0.91$). The deviating cell is **F7**, $C+U-$.

D.4 C-OPERATIONALIZATION ROBUSTNESS

The C verdict requires a rule for combining the two passive endpoints and for treating degenerate cells. The frozen primary rule is the disjunction of Appendix B.4. The full matrix was mechanically recounted under two alternatives fixed at recount time, a stricter conjunction (authority alignment required at *both* endpoints) and the primary rule with U-degenerate cells excluded rather than scored.

rule	C^+U^+	C^-U^+	C^+U^-	C^-U^-	indeterminate
disjunction (primary)	28	7	3	4	—
conjunction	24	11	1	6	—
disjunction, degenerate-excluded	28	5	2	2	5

All four regimes exist under every rule. The exact counts are rule-dependent and are therefore not interpreted as prevalence estimates or rule-invariant quantities. Eight cells are threshold-adjacent, meaning a one-episode change could flip the verdict. They are M1-F7, M3-F1, M3-F4, M3-F6, M4-F4, M4-F7, M5-F5, and M5-F7. The recount reproduces the frozen per-cell verdicts on all 37 cells with continuous records. M1's five earliest cells carry boolean-only records and are flagged as such.

Calibration-tolerant recount. A separate recount applies the two secondary rules of Appendix B.5, which hold the fitted direction and polarity frozen and vary only how the transported readout is scored. This analysis requires per-cell transported scores at the evaluation endpoints. Full transported scores at both endpoints were recoverable for 36 of the 42 cells. Inclusion required reproduction of the archived primary statistic under the prespecified tolerance ($|\Delta\text{acc}| \leq 0.02$), and the six remaining cells (M3-F1, M4-F1, M6-F1, M1-F3, M1-F4, M1-F5) are left unmeasured in this secondary recount because their archived statistics cannot be mechanically reproduced from retained artifacts under that tolerance.

rule (36 measured cells)	C^+U^+	C^-U^+	C^+U^-	C^-U^-	unmeasured (of 42)
primary (frozen threshold)	26	5	3	2	6
AUROC-C	29	3	3	1	6
Intercept-C	28	4	3	1	6

All four C/U regimes remain represented under both calibration-tolerant rules on the measured full-resolution cells. Six cells change verdict relative to the primary rule. M1-F1 becomes positive under both rules (AUROC $0.861/0.902$ with clustered CI clear of chance, cross-validated intercept-recalibrated BA 0.740 $[0.67, 0.81]$ / 0.771 $[0.68, 0.85]$). M3-F6 and M4-F7 become positive under both rules. M5-F3 becomes positive under Intercept-C only (0.722 $[0.58, 0.86]$), M5-F5 under AUROC-C only (0.674 $[0.59, 0.76]$), and M3-F3 loses its positive verdict under Intercept-C only (a marginal CI miss). The remaining cells keep their primary verdicts, and the stable C^+ cells are ranking-robust (AUROC typically at or near 1.0). As with the endpoint recount, exact counts are not prevalence estimates.

The recount shows that the same primary C^- verdict can arise from qualitatively different transported-readout failures, whether preserved ranking with failed calibration, mixed transport degradation, or failure or inversion of the transported ranking itself. Three cells span this range. M1-F1 is *calibration-limited*. The authoritative-value ranking survives (0.86–0.90 AUROC) while the frozen decision boundary fails. M4-F7 is a near-pure calibration failure. The archived frozen-threshold accuracy sits at chance (0.500 at both endpoints) while the transported ranking is perfect (AUROC 1.0 at both). M1-F7 is the strongest *ranking-level* failure in the matrix. AUROC is exactly 0.0 at both endpoints under the frozen polarity, so the transported readout ranks against the authoritative value. M2-F1 is a further ranking-level case (AUROC 0.374, consistent with the order-confounded refit of Appendix D.5), and the C^-U^- regime survives through M4-F4, whose profile is mixed-intermediate (AUROC 0.38–0.49 with partial intercept recovery). These profiles are diagnostically distinct, and per-cell classifications appear in Appendix D.5.

Measured cells use the surviving serialized decoders where these are retained. The twelve F6 and F7 cells use decoders deterministically reconstructed from the frozen agreement activations under

the archived fitting protocol, and all twelve reproduce their archived P0/P1 primary statistics exactly before any secondary scoring. No decoder was tuned, and cells whose scores could not be recovered are left unmeasured rather than filled by inference.

D.5 READOUT-FAILURE DIAGNOSTICS

Agreement-trained readouts fail in different ways under conflict. Some retain AUTH ranking but lose calibration, while others lose or invert the transported ranking. For each C^- or borderline cell, three diagnostics were computed on the conflict cells: the transported decoder's AUROC (ranking quality independent of threshold), the accuracy of an intercept-only recalibration (threshold refit on conflict cells, direction frozen), and the accuracy of a full refit within the conflict context (Appendix C.3). The classification rule was fixed before the cross-model measurements. *Ranking-level* means low transported AUROC, failed intercept-only recalibration, and a successful conflict refit. *Calibration-level* means high transported AUROC with intercept-only recalibration alone recovering accuracy (direction intact). *Mixed-intermediate* means partial features of both. We additionally use *calibration-limited* for the case in which substantial ranking ability remains and intercept-only recalibration recovers most of the classification loss, despite some degradation in ranking quality. These labels describe different ways in which the agreement-trained readout fails under conflict. They are not established mechanisms.

cell	profile	AUROC (transported)	intercept-only	conflict refit	mode
M1-F1	$C-U+$	0.861	0.740 (CV)	0.896	calibration-limited
M2-F1†	$C-U+$	0.374	fails	0.942	ranking-level
M3-F1	$C-U-$	0.549	0.543	0.894	ranking-level
M1-F3	$C-U+$	low	fails	0.722	ranking-level
M5-F3	$C-U+$	0.67/0.76	partial	0.75–0.78	mixed
M5-F5	$C-U+$	0.503	0.500	1.000	ranking-level
M1-F5	$C-U+$	**0.879**	**0.845**	0.991	calibration-level
M3-F6	$C-U-$	high	recovers	—	calibration-level
M4-F7	$C-U-$	1.000	recovers	—	calibration-level
M6-F3	borderline $C+$	0.96/0.91	0.94/0.86	—	calibration-level
M4-F4	$C-U-$	0.618 (pred. frac. 0.00)	0.684	1.000	mixed-intermediate
M1-F7	$C-U+$	0.000	sign-refit 1.000	—	polarity inversion

†M2-F1's conflict-context refit is order-confounded. A balanced factorial test that crosses the authoritative and last-mentioned values independently (144 rows / 36 repositories) shows the conflict-fitted readout places more weight on recency than on authority. Its recency coefficient is -0.60 to -0.63, about $2.7\times$ the authoritative component at the position nearest the decision, and its agreement with the authoritative value is 1.00 under the presentation order it was fitted on but 0.49 under the reversed order. All standard-order M2 refit figures should therefore be read as order-specific rather than as order-invariant recovery of an authority-tracking readout.

Every ranking-level case retains strongly decodable authoritative-state information within the conflict context while the transported agreement-trained readout fails to rank AUTH reliably. Every C^+ control cell shows within-conflict accuracy $\geq$ transported accuracy, providing a consistency check for the diagnostic. Among the profiled C^- and borderline cells, ranking-level failure is confirmed on M2-F1, M3-F1, M1-F3, and M5-F5. Polarity inversion (a fitted direction whose sign, not existence, fails to transport) occurs once, on M1-F7. Under crossed fitting the same cell reaches AUROC 1.000 (Appendix C.6), so the anti-AUTH ranking characterizes the agreement-trained readout rather than an absence of recoverable information about the current value. The primary M1-F1 cell is calibration-limited with a reduction in ranking quality (AUROC 1.000 under agreement, 0.861 under conflict, cross-validated intercept-only recalibration 0.740 $[0.67, 0.81]$). The geometry of its agreement-trained readout is analyzed in Appendix C.4, and the crossed-fit W readout for the same cell transports without recalibration (Appendix C.5). M2-F1's transported AUROC of 0.374 sits below chance with direction and polarity frozen, consistent with its order-confounded refit, and the same cell reaches crossed-fit balanced accuracy 0.842/0.789 (Appendix C.6). M1-F5's transported accuracy sits at the classification boundary (lower bound exactly 0.500) with the largest within-versus-transport gap of the profiled cases (0.991 vs. ~ 0.50). Its diagnostics resolve it as calibration-level, so it is not grouped with the ranking-level failures.

D.6 TARGETED FAMILY CONTROLS

Outcome-token neutralization (F4-hard). Because several stable families textually contain outcome words, a hardened F4 variant removes every explicit outcome token (`PASSED`/`FAILED`/`ERROR` lines) from the execution evidence while preserving tracebacks and assertion output, holding the rest of the cell construction fixed. M1's pre-demand passive AUTH accuracy is 0.8947 under both native and neutralized evidence (numerically unchanged), and M3 remains C^+ under neutralization (0.789 at the second endpoint). The M5 arm preserves its native C statistics but carries a degenerate action gate on both arms, so it contributes the C-preservation observation only, and the M4 arm yields no measurement, because its action policy is constant under the neutralized evidence. Stable F4 currentness therefore does not require direct outcome-label tokens. The control removes one candidate carrier. It does not causally identify tracebacks or assertions as the carrier.

Patch applicability (F6). F6's apply-check evidence contains explicit outcome-like wording ("applies cleanly" vs. "does NOT apply"), placing it in the same explicit-evidence regime as F2. F6 is C^+U^+ on five of six models (at or near ceiling on M1, stable on M2, M4, M5, and M6). M3-F6 is the exception and exhibits a calibration-level failure, showing that even explicit-evidence families are not unconditionally stable across models.

Executed provenance (F7). F7 is retained as a stress family. Its label requires provenance inference rather than evidence transcription, and F7 spans all four C/U regimes across the six models, including both dissociation directions. The results support no single-mechanism summary of F7. Its role in the paper is to bound generalization claims about families of this type.

E CAUSAL CONTROL AND PERSISTENCE

E.1 INTERVENTION CONSTRUCTIONS AND VALIDITY CRITERIA

Two different intervention constructions are used, and they are not interchangeable.

(i) Opposite-condition transplantation (F1). For a given row we take the hidden state of the *same* episode rendered under the opposite authoritative condition and write that full vector into the receiver's residual stream at one layer and one position (pre-norm, exact transplant). This replaces the entire representation at that site rather than adding a calibrated direction, so it carries a substantial nonspecific disruption component. The matched **SAME** arm (a transplant from a same-condition donor) measures that component.

(ii) Calibrated class-contrast writes (M5-F2, M5-F5). Class-conditional means of the layer-ℓ residual stream are computed on held-out calibration episodes, and their difference u is added with magnitude $m \cdot u$, where $m = 1$ is the natural class displacement. In the F2 construction the state and answer contrasts coincide, so the effective direction is not uniquely identified as state-specific. Three controls are used. **ORTH** consists of norm-matched directions drawn orthogonal to u (20 seeds per row, seeds disjoint across experiments) and is the primary specificity control. **SAME** is a same-class donor write, auxiliary only, because in the original commitment-alignment battery its magnitude parity was not recoverable, so no specificity claim relies on it. **REVERSE** is a sign-inverted write used only as a monotonicity check and excluded from specificity tallies.

Three outcomes are kept separate throughout. They are movement of an *internal score* (a teacher-forced candidate-token margin), change of the *emitted action token* at the pass where the model chooses its action, and change of the *final action* that the frozen parser reads at the end of the trajectory. An intervention can achieve the first without the second, and the second without the third.

Positions matter as much as directions. The *prompt-final* position is the last position of the rendered prompt, where a teacher-forced margin is conventionally scored. The *action-producing pass* is the forward pass whose next-token distribution actually selects between the two candidate actions. In the primary model this occurs hundreds of tokens into generation, after a reasoning span. A *renewed action decision* is a later pass at which the model again selects between those candidates. We use the shorthand *deciding pass* only for the action-producing pass.

Pre-specification status is stated with each result below (pre-specified in a frozen protocol before outcome inspection, held-out replication, or outcome-seen diagnostic). The frozen protocol records are archived with the experimental series under fixed identifiers (M1-F1 single write `66c59f50`, clamped persistence `03ef0b56`, M2-F1 confirmatory `2ab20d33`). Eligibility rules define cohorts before intervention and are never outcome-selected. All M1-F1 results use an audited event mapping, and earlier detector-driven runs that failed the audit are excluded.

E.2 M1-F1 INTERVENTION AT THE ACTION-PRODUCING PASS

This pre-specified experiment uses the corrected event mapping, with candidate sets restricted to true action first-tokens, and the exact pre-norm transplant convention. The event mapping was verified on all 14 audited rows, and hook-versus-stored and patch-equality checks were exactly zero. The cohort comprises 30 primary rows over 17 repositories, with a single write ($K{=}1$) at one layer at the true action-producing pass.

arm	emitted token changed	opposite-action token	final action reversed
TARGET	24/30	**13/30**	0/30
SAME	20/30	5/30	0/30

The late representation has condition-specific causal influence on the emitted action. TARGET pushes the token to the opposite action on 13/30 rows against 5/30 for SAME. The full-vector transplant is also substantially disruptive in both arms (24/30 and 20/30 tokens changed), so the condition-specific component is real but modest, not a clean flip.

The causal chain therefore reaches the emitted action token. It does not reach the final action, which is unchanged in both arms (0/30). Because the emitted token did change on 13 rows, this null is measured rather than inferred, and it locates the break between the flipped token and the end of the trajectory. What happens after the flip is the subject of Appendix E.5.

E.3 M5-F2 DECISION-POSITION ALIGNMENT

This experiment tests whether behavioral causal effects depend on intervention position. In a requirement-satisfaction cell, a calibrated class-contrast write strongly moves the teacher-forced candidate-token margin.

At the prompt-final position (26 eligible rows / 16 repositories; calibration-derived scale $q = 5.36$) the write crosses the margin's sign boundary on 16/26 rows; SAME and an order-only control cross 0. The committed answer changes on 0/26 (ORTH free-generation 0/78). The behavioral endpoint is independently movable on this construction, because changing the underlying evidence moves the committed answer in 16/16 paired checks. A liveness audit explains the mismatch. At the prompt-final position the model's next token is a formatting token, and the scored answer alternatives are not the live next-token alternatives there. The site fails this liveness check in all 48 tested cases across three standard prompt formats (answer-suffix $0/16$, ReAct $0/16$, JSON tool-call $0/16$), so the failure is not specific to our template among the formats tested.

At the action-producing pass, the same calibrated direction at the same magnitude changes the committed answer on 13/26 of the original rows. Applying the write persistently at every subsequent decode step does not improve on applying it only at the deciding pass. The two agree row-wise, so position alignment rather than persistence is the operative factor. A substitution-scale diagnostic bounds the write. It moves the margin by a median $0.375\times$ the natural class gap (range 0.23–0.51, norm ratio 0.17–$0.18\times$ the deciding-position residual norm), below the $\approx 1.0\times$ signature of output substitution.

Pre-specified held-out replication. On 14 intervention-naive episodes (28 rows), with the calibration direction reused refit-free and an in-run unpatched baseline, 14/14 eligible UNSAT→SAT cases change, 0/14 SAT→UNSAT, REVERSE 0/28, SAME 0/28, and ORTH $\times 20$ 0/560. Because the calibration episodes are shared with the original run, this is a held-out *sample* replication, not an independent replication. Combined across the two F2 cohorts, 27/29 eligible UNSAT commitments change, 0/25 SAT commitments change, and orthogonal controls give 0/520 (26×20 seeds) plus

0/560 (28×20 seeds), i.e. 0/1080 within F2. The M5-F5 battery below and the dose-response battery are reported separately and never pooled into this figure.

The conceptual result is that a crossed internal candidate score is not sufficient for a behavioral causal claim if the intervention is evaluated at the wrong computational position.

Pre-specified SAT dose response. To test the direction asymmetry at $m{=}1$, the SAT-direction write was applied at increasing magnitude to all 30 baseline-SAT rows drawn from the two F2 pools, selected only by in-run baseline class and all retained. Because those pools' outcomes had been inspected, this is an outcome-seen diagnostic cohort, not an independent replication.

dose	median Δmargin (nats)	crossings	SAT$\rightarrow$UNSAT	subst. scale
$m{=}1$	-4.50 (base $+8.75$)	0/30	0/30	0.25
$m{=}2$	-11.69	30/30	30/30	0.65
$m{=}4$	-15.63	30/30	30/30	0.86

Orthogonal controls give 0/600 crossings and 0/600 reversals at each of $m{=}2$ and $m{=}4$ (0/1200 across the diagnostic). Rowwise dose monotonicity holds on 30/30 rows, with $P(\text{reversal} \mid \text{crossing}) = 1.00$ at $m{=}2$ and $m{=}4$ and $P(\text{reversal} \mid \text{no crossing}) = 0$ at $m{=}1$. The $m{=}1$ asymmetry is thus distance-consistent, because SAT rows start farther from the deciding boundary and do not cross at natural magnitude. Both elevated doses exceed the frozen sub-substitution-scale bound and are diagnostic only. They are not evidence of symmetric natural-scale causal control.

E.4 M5-F5 RUNTIME-RESOLUTION SCOPE CONTROL

This pre-specified test bounds the F2 result. In the runtime-resolution family, independently calibrated state-contrast and answer-contrast directions are close to orthogonal ($\cos(u_{\text{state}}, u_{\text{ans}}) = -0.026$; $\|u_{\text{state}}\| = 39.05$, $\|u_{\text{ans}}\| = 12.25$, with a norm-matched answer arm). The margin is scored at the first divergent answer token and the cohort is boundary-screened (26 smallest-margin rows, median $|\text{margin}| = 3.375$ nats against a frozen threshold of 3.74). The transferred power criterion was satisfied, but because the threshold was calibrated on F2 and cross-family margin-scale comparability is not established, we treat that behavioral-power claim as conditional.

arm (26 rows each)	reversals	mean $\lvert$margin shift$\rvert$ (nats)
STATE	0/26	0.043
ANSWER	0/26	0.034
ANSWER norm-matched	0/26	$\sim$0.05
ORTH $\times 20$	0/520	median 0.0

In F2 the same construction at comparable norm ratio moves the deciding-pass margin by 4.5–7.0 nats; here it moves it by about 0.04 nats, indistinguishable from the orthogonal floor. Because the margin itself did not move, no reading in which a crossed margin was behaviorally ignored is available. The result bounds the scope of the F2 finding. The tested class-contrast directions do not produce the F2-like behavioral effect in this family, so the F2 effect does not extend automatically across task-state families under the same kind of intervention.

E.5 M1-F1 TRAJECTORY PERSISTENCE AND SELF-REPAIR

This pre-specified test asks what happens after the action token is flipped. It uses the corrected detector and adds a clamped arm that re-applies the transplant at up to four genuine action decisions. The cohort comprises 26 primary rows over 17 repositories.

arm	opposite-action token	final reversed	median decisions
TARGET, single event	10/26	0/26	1
TARGET, clamped over 4 events	**19/26**	1/26	4
SAME, clamped over 4 events	13/26	0/26	4

Clamping raises the rate at which an opposite action is emitted (10/26 to 19/26) but barely moves the final action (1/26, against 0/26 for SAME), so the single clamp success is not clearly beyond the disruption floor.

The event traces directly show the sequence. An opposite action is emitted, a later renewed action decision occurs, and the original action is re-established. Recovery is fast in the single-event arm (observed latencies from 1 token, median around 5, with several rows re-deciding at the very next action event) and slower under clamping, where the observed range extends to 29 tokens because each re-intervention re-flips the token. The model nevertheless outlasts the four-event budget on 25/26 rows: once the clamp is exhausted, the next unpatched decision restores the baseline action. Repeated intervention therefore delays recovery but usually does not prevent it. We refer to the observed flip–redecision–restoration sequence as behavioral self-repair.

The renewed decisions are re-decisions over a highly transportable decode-time signal rather than evidence of repeated re-encoding of the state. A powered transport study over eight recorded positions (79 rows / 50 repositories; all positions retained; local diagonals ≥ 0.944) finds that probes fitted at one decision event evaluate at the others near ceiling, with median $R \approx 0.99$ within the decision-event block and AUROC 1.00 from every decision event to the action-string onset. In this pre-specified test, the renewed decision positions share a highly transportable decode-time readout of the authoritative value. The observed transport pattern is inconsistent with repeated event-specific re-encoding as operationalized here, which predicts high diagonals with low off-diagonal transport. The transport losses that do exist are frame-specific, not event-specific, and concern the prompt-final position. A label-free rank-≤ 32 map fitted on paired states recovers 73% of the prompt-final-to-decode gap and generalizes across later decision events (median 0.73), while a shuffled-pair map is destructive, consistent with a systematic frame difference between the prompt-final and decode-time positions rather than a population offset. This frame difference is model-specific. The same comparison gives a deficit of 0.027 in M2 against 0.390 in M1, so it is an implementation property and not an explanation of the C^- regime. Full matrices and the alignment analysis are in Appendix C.9.

E.6 Cross-model persistence comparison

M4-F1. In the powered test at the text-derived action position ($K{=}1$), opposite-condition transplantation changes the emitted action token on 18/18 rows and the final action on 8–10/18 depending on the layer arm, with SAME at 0/18 (paired difference $+0.556$ $[0.389, 0.737]$). A traced replay of this configuration with the corrected detector (28 traced rows, of which 12 show a target-directed immediate token change) records the temporal structure directly. 10/12 changes persist to the final action with no renewed decision event on those rows, while the 2/12 rows that restored the original action each contain one renewed decision event (at latencies of 14 and 17 tokens). In this traced sample, both recoveries coincide with a later decision opportunity, whereas all ten persistent changes occur without one, and such opportunities are rare in M4.

M2-F1. The confirmatory cohort was pre-specified, and rows were eligible only when the baseline action was correct (40 eligible armed rows / 36 repositories; 44 ineligible rows screened out before intervention). TARGET changes the immediate action target-directedly on 16/40 against SAME 6/40, meeting the prespecified criterion for a target-specific causal effect. All 16 target-directed changes remain the final action, and renewed decision events number 0 on every one of those rows. Under the native F1 protocol M2 answers in about three tokens with no reasoning span, so its answer terminates after the initial decision and the trajectory provides no later decision opportunity. This is a structural property of the native dynamics and leaves recovery capacity untested.

Descriptively, across the three models recovery is observed when a later decision opportunity occurs. M1 has many such events (median 4), M4 has very few, and M2 has none under the native F1 protocol. We do not attribute this difference to the authoritative-currentness class. The same C/U regime is multiply realized, and these experiments do not establish a causal link from currentness class to temporal decision organization.

E.7 State versus answer-token identity (factorial intervention)

A separate factorial experiment disentangles the task-state code from the answer-token code. In M1, a 2×2 design crosses the underlying state with the mapping between states and answer to-

kens, so that the two factors vary independently (88 episodes, 352 rows). Behavior is symmetric across the two mappings, so answer-token identity carries no behavioral privilege of its own. In this factorial construction the two codes are identifiable separately. At L31, the state direction reaches cross-validated AUROC 0.879 and the answer direction 0.987, each estimated with the other factor balanced, and cross-decoding confirms that each direction carries its own factor. The two directions are nearly orthogonal (cosine -0.045).

The causal test applies natural-scale displacements ($m{=}1$) at the live answer-producing pass ($K{=}1$) on 85 eligible rows (44/41 across the two states), with the answer direction as a matched comparator and eight norm-matched orthogonal controls. The pre-registered behavioral success criterion fails. No arm produces a margin crossing, an emitted-answer change, or a final-answer change in any of the 680 row-arms. The informative effect appears in the margin shift. The state direction moves the answer margin by a median 6.38 logits toward the opposite-state answer in 75/85 rows, while the answer direction moves it -0.12 and the orthogonal controls 0.25. The push is roughly $2.6\times$ below the median committed margin, so the leverage is direction-specific but below the flip threshold at this scale. At this scale, the more decodable answer direction has negligible influence on the margin, while the less decodable state direction produces a substantial shift. At the matched natural intervention scale tested here, greater decodability does not imply greater influence on the live decision margin.

E.8 Summary of causal evidence

The causal experiments distinguish four outcomes, each established by specific cells.

1. *An internal quantity moves.* Established by the M5-F2 prompt-final test (16/26 margin sign crossings) and by the dose response.
2. *The intervention reaches the action-producing pass.* Established by M1-F1 (13/30 opposite-action tokens), M5-F2 at the deciding pass, M4-F1 (18/18) and M2-F1 (16/40). Not established by the M5-F2 prompt-final site, which fails the liveness check in all 48 tested cases.
3. *The emitted action changes.* Established by M1-F1, M4-F1, M2-F1 and M5-F2 (13/26). Not established by M5-F5, where the margin itself does not move.
4. *The change survives to the final action.* Established by M4-F1 (8–10/18) and M2-F1 (16/16 of the target-directed flips). Not established by M1-F1, where the final action is unchanged (0/30 single write; 1/26 under a four-decision clamp).

Together, the four levels separate what an intervention moves internally from what it does to the action being produced, and what it does to the action from what survives the rest of the trajectory. No result at any level is attributed to the C^+/C^- class.