ACTIVE NEURAL MATTER: FINITE-WINDOW RESPONSE IN PHYSICAL SYSTEMS AND AI WORKFLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

We formulate Active Neural Matter (ANM), a finite-window response framework spanning physical systems and distributed AI workflows. Its three elements are controlled source perturbations, a candidate retained state representing information carried between interactions, and scalar or vector readouts (observable outputs). Separate source-to-state and source-to-readout measurements connect forward analysis of propagation and readout sensitivity with inverse tests of what information a representation must preserve, including relevant history. Representation choice becomes an experimental hypothesis: for specified perturbations, readouts and windows, output differences beyond measurement error at identical retained states reject adequacy. Approximate matches test readout models with independently justified sensitivity bounds; informative comparisons and error controls distinguish rejection, local support within stated tolerances and unresolved adequacy. The framework covers continuous and discrete dynamics, finite contrasts without assuming derivatives, and optional nested boundaries for continued operation and stricter correctness. Delayed consensus calibrates measured responses against analytic predictions, while a polar flock compares nonlinear steady-state readouts. Recorded-program comparisons and a constructed omitted-input control reject inadequate representations; an exploratory constructed tool workflow restores a task-required response on one known case by exposing existing request–return associations without adding answers. ANM links response prediction to testable requirements on retained information, providing a common experimental basis for refining representations and studying information transfer and consumer use.

1 INTRODUCTION

An AI workflow may preserve a tool result yet lose the request that gives it meaning; it may also preserve both records and interpret their association incorrectly. The same wrong answer can therefore call for different changes to information retention or downstream use. Delayed physical networks pose a related measurement problem: the same current field can evolve differently because of its history. Both settings require examining information retained between interactions. We call a declared field, graph coordinate or history-dependent representation of that information a *retained state*. The endpoint readout is the measured output at the end of the observation window. Comparing state and readout changes can distinguish propagation from readout sensitivity within a specified model (Figure 1).

Active Neural Matter (ANM) makes this response relation the unit of analysis: a controlled change to an input (the source perturbation), a candidate retained-state displacement and the readout change over a finite window. *The choice of representation itself becomes an experimental hypothesis.* Forward analysis traces responses for specified states and dynamics; inverse analysis asks which distinctions a representation must preserve. With the window and other conditions fixed, identical retained states with resolved readout differences identify histories a revised representation must separate. They do not uniquely identify a missing variable. Approximate matches require a specified class of readout models with an independently justified bound on sensitivity—output change per unit state change—and controlled measurement errors (Section 2).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

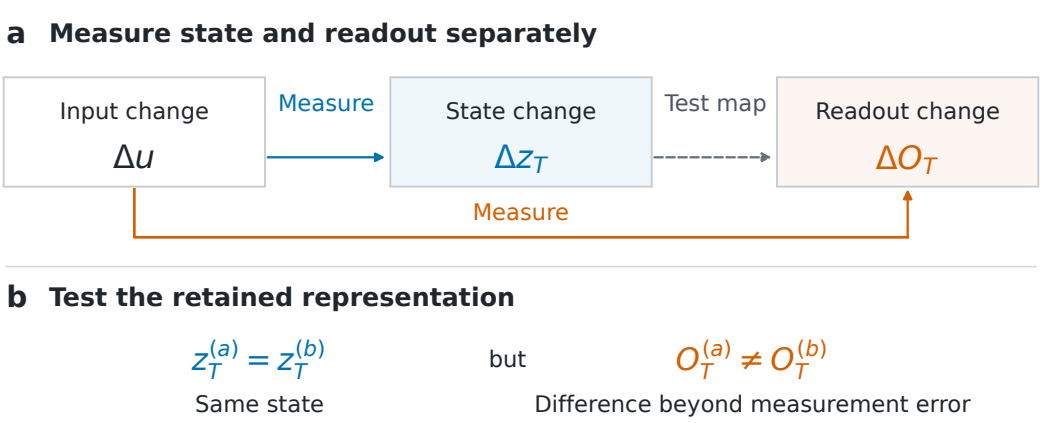


Figure 1: **Forward response and inverse representation tests.** **a**, A finite source contrast Δu produces separately measured state and scalar or vector readout changes $\Delta z_T, \Delta O_T$. Solid arrows denote measurements; the dashed state-to-readout relation is a hypothesis, not an established causal path. **b**, Histories a, b with identical states but readout differences beyond measurement error must be distinguished by an adequate representation. Subsequent inputs are fixed; approximate matches require sensitivity and error bounds (Section 2).

In distributed AI workflows, producing components supply sources, communication records retain information, and consuming components produce readouts. These separate roles and interfaces permit audits without inspecting model weights. An externally specified graph and delayed-tool history instantiate them; jointly formed live multi-agent states require separate experiments.

Relation to established methods. Response theory connects perturbations to observables (Kubo, 1957; 1966; Ruelle, 2009), and projection methods explain why reduced variables retain history (Mori, 1965; Zwanzig, 1961). Diagnosis, causal abstraction and functional observability test mechanisms or information preserved by a representation (Gertler, 1991; Geiger et al., 2025; Montanari et al., 2022a). Activation interventions and mediation analysis test specified internal pathways (Zhang & Nanda, 2024; Syed et al., 2024; Vig et al., 2020); workflow methods examine fault locations, handoff preservation and memory pathways (Zhang et al., 2025; Ma et al., 2025; Sun et al., 2026; Xiao et al., 2026). ANM combines these operations in an experimental specification: declare the source class, retained state, readout and window; measure state and endpoint responses separately; and check whether the comparisons can test the representation. The resulting judgment guides the next measurement or a revised representation, with operating criteria defined by the readouts. Appendix A compares access requirements and conclusions with these methods and collective-dynamics and graph approaches.

Contributions. *(1) Separate propagation from readout sensitivity.* Across physical systems and AI workflows, we specify paired state and endpoint measurements over a finite window, including the histories needed to interpret them. Within stated models, these measurements distinguish propagation and readout changes that produce identical endpoint responses. *(2) Test and guide refinement of the representation.* First-order and finite-contrast tests identify distinctions a candidate must retain; design conditions distinguish rejection, bounded local support and unresolved adequacy. *(3) Connect responses to operating criteria.* We link general scalar or vector readouts to operating boundaries. One special case nests a stricter correctness criterion (exactness) inside continued operation (workability). Together, they make representation choice part of experimental design: what to perturb, measure and retain, and which comparison to perform next.

Delayed consensus, flock and AI workflows supply complementary calibration, readout and diagnostic evidence (Appendix Table 7).

2 SOURCE, RETAINED STATE AND OBSERVABLE RESPONSE

A delayed system can have the same instantaneous field under two histories but different later endpoints under identical permitted inputs. An adequate retained representation must distinguish the relevant histories; it need not preserve a particular raw record or be unique. We distinguish a shared field c_t, the full memory-augmented state s_t, and a candidate retained coordinate $z_t = G_f^{\mathrm{cg}}(s_t)$ for measurement context f. The coarse-graining map $G_f^{\mathrm{cg}} : X_f \to Z_f$ maps augmented-state space X_f to retained-coordinate space Z_f. Its choice does not establish sufficiency or autonomous dynamics. In delayed consensus, the relevant history is the recent physical field; in a tool workflow it may include requests that explain later returns. The common specification fixes their measurement roles, while their coordinates and dynamics remain realization-specific.

Response over a finite observation window. Include required memories, histories, readout inputs and observation records in s_t. Let u_t collect the controlled source coordinates. The readout $O_T(s_T) = (O_{1,T}, \dots, O_{d,T}) \in \mathbb{R}^d$ collects $d \geq 1$ measured observables at time T. Window summaries such as amplitude and phase require the accumulated measurements in s_T (Appendix B.2). Fix preparation and nonperturbed inputs. We use Δ for finite contrasts between controlled inputs and their responses, and δ for tangent perturbations. On a declared time grid write the transition as $s_t = F_t(s_{t-1}, u_t)$, with $A_t = \mathrm{D}_s F_t$ and $B_t = \mathrm{D}_u F_t$ evaluated along the baseline trajectory. Differentiability of the transition, coarse-graining and readout maps on the perturbation space gives

$$\delta s_t = A_t \delta s_{t-1} + B_t \delta u_t, \tag{1}$$

$$\delta O_T = \mathrm{D}_s O_T(s_T) \sum_{t=1}^{T} \Phi(T,t) B_t \delta u_t, \qquad \Phi(T,t) = A_T \cdots A_{t+1}. \tag{2}$$

Each term follows one perturbation: B_t injects it, successive transitions A_t propagate it, and $\mathrm{D}_s O_T$ converts the terminal displacement into a measured change. Here $\Phi(T,T) = I$. The formula fixes the initial state; varying its history adds $(\mathrm{D}_s O_T)\Phi(T,0)\delta s_0$. If the readout factors differentiably as $O_T = \bar{O}_T \circ G_f^{\mathrm{cg}}$, its differential is $\mathrm{D}_s O_T = \mathrm{D}_z \bar{O}_T \, \mathrm{D}_s G_f^{\mathrm{cg}}$. This is the candidate composition to test. The coarse-graining Jacobian is the identity only when the graph itself contains the required memory and serves as the retained coordinate. Derivatives require more than Lipschitz well-posedness; nonsmooth points require finite contrasts or specified one-sided responses. We compare the composed and independently measured endpoint responses. *Step refinement* compares response slopes at a perturbation size and a smaller size to assess a first-order interpretation. Figure 2 shows why state measurements and selective state interventions answer different questions.

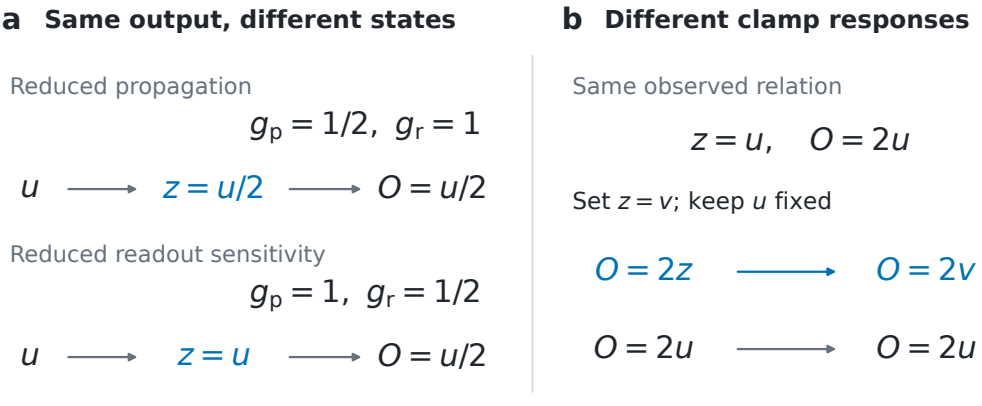


Figure 2: **State measurement and causal intervention answer different questions.** **a**, In $z = g_p u$, $O = g_r z$, g_p is the propagation gain and g_r the readout gain. Halving either from $g_p = g_r = 1$ gives the same endpoint $O = u/2$, but different states. Restoring the altered gain restores $O = u$; ordinary state-space identification gives the same decision. **b**, The models $O = 2z$ and $O = 2u$, both with $z = u$, have identical source-response records. Independently setting $z = v$ while holding u fixed (a clamp) separates them. These are analytic models, not AI measurements (Appendix B.2.4).

A falsifiable retained-state test. On a finite-dimensional measured source space $\mathcal{U}_T$, let $\mathcal{K}_{z,T}$ and $\mathcal{K}_{O,T}$ be the first-order source-to-state and source-to-endpoint maps. Existence of a linear readout map is characterized by

$$\left[\exists\, \mathcal{H}_T^{(1)} :\, \mathcal{K}_{O,T} = \mathcal{H}_T^{(1)} \mathcal{K}_{z,T}\right] \quad \Longleftrightarrow \quad \ker \mathcal{K}_{z,T} \subseteq \ker \mathcal{K}_{O,T}. \tag{3}$$

The map is unique on $\operatorname{Im} \mathcal{K}_{z,T}$, the reached state displacements. The kernel contains perturbations with no first-order state response: if the readout responds to one, no linear readout of that state can explain it. Such a null direction requires $\dim \mathcal{U}_T > \operatorname{rank} \mathcal{K}_{z,T}$; with full column rank the relation is automatic. For C^1 source-to-state and source-to-endpoint maps, constant rank of the former and kernel inclusion throughout a neighborhood give local nonlinear factorization. Pointwise kernel inclusion alone does not establish nonlinear factorization: $z(u) = u^2$, $O(u) = u^3$ obey the derivative criterion but identify opposite inputs with different endpoints. Exact neighborhood sufficiency instead requires equal retained states to imply equal endpoints; stochastic readouts require repeated controlled continuations and an appropriate distributional comparison. The proof, delayed-history counterexample and analytic diagnostic controls appear in Appendix B.2.4.

Apply the criterion after checking response resolution and, for a first-order claim, step refinement. Then check whether the perturbations include a resolved direction in the state-response kernel or a controlled same-state comparison. A measured endpoint difference exceeding the justified state-matching and noise allowance rejects the bounded readout class; local support requires adequate null-direction coverage, state-matching error controlled by a justified readout-sensitivity bound, endpoint-contrast intervals within prespecified equivalence bounds and a detectable positive control. Unresolved comparisons remain undecided. These judgments concern particular directions and readouts; they need not be mutually exclusive across an experiment.

Physical realization and conditional limits. A controlled semilinear field model connects this protocol to neural-field relaxation, local interaction and delayed inputs (Wilson & Cowan, 1972; Amari, 1977; Bressloff, 2012):

$$\partial_t c_t = L c_t + \sum_{i=1}^{N} J_i(c_t, m_i(t)) + \sum_{\ell=1}^{N_\tau} E_\ell y_\ell(t - \tau_\ell). \tag{4}$$

Here $c_t \in L^2(X, \mathbb{R}^q)$ has q channels on a spatial or graph domain X; L generates relaxation, J_i is a local source with memory m_i, and E_ℓ injects a prescribed signal y_ℓ after delay τ_ℓ. Internal feedback histories, suppressed in the displayed source arguments, enter the augmented state. Appendix B.2 specifies the read/update/write realization of J_i; Appendix B.3 gives existence conditions. Autonomous reduction and steady susceptibilities require additional closure and stability assumptions (Appendix B.2.1). Finite-window measurements require neither limit; the field model supplies a realization, not a shared microscopic law.

3 Readout-defined operating boundaries

The general protocol allows one or more observables. This two-component example separates continued operation from stricter endpoint correctness. Fix a family $f = (\Theta_f, \Lambda_f, b_f, P_f, Q_f)$: instance class Θ_f, interface and stopping rule Λ_f, loading coordinate b_f, and two readouts on the same population or protocol. The stopping rule fixes when the endpoint is evaluated. Workability P_f measures continued admissibility until that endpoint; exactness Q_f additionally requires its correctness criterion. The pair $O = (P_f, Q_f)$ must satisfy $0 \le Q_f \le P_f \le 1$ by logical inclusion, not a fit to observed values. Define

$$w = P_f, \qquad r = Q_f / P_f \quad \text{only where } P_f > 0. \tag{5}$$

Here w is defined throughout the domain. The ratio is conditional exactness only for nested measures on the same population, and otherwise is an inner-to-outer ratio. The four regimes are exact conditional on workability ($w > 0,\ r = 1$), partially exact ($w > 0,\ 0 < r < 1$), workable but inexact ($w > 0,\ r = 0$), and failed ($w = 0$, with r undefined). At $r = 1$ the two readouts are equal; w separately records workability. Measured boundary assignments require stated uncertainty and tolerances. Readout level sets define boundaries; dynamics determine how a loading path approaches them. Encountering all four regimes in order requires additional regularity, monotonicity

and separated-crossing assumptions. Coincident thresholds can skip regimes, and nonmonotone recovery remains compatible with nested readouts (Appendix D.1).

Source perturbations can change either readout; their responses are measured separately. Binary indicators require finite contrasts or stated smooth extensions. The readout pair determines neither state dimension nor sufficiency, and response magnitude alone does not measure intervention cost.

4 Complementary physical realizations

Delayed consensus calibrates a driven finite-window response, whereas a polar flock separates nonlinear steady-state readout scales. Standard delay-system and active-matter descriptions supply their respective reference models (Hale & Verduyn Lunel, 1993; Olfati-Saber & Murray, 2004; Vicsek et al., 1995; Toner & Tu, 1995; Marchetti et al., 2013). The shared measurement procedure assigns a distinct role to each system; their dynamics and response functions remain realization-specific.

Driven delayed consensus. A spatially uniform sinusoidal source drives a degree-four ring of 64 noisy dissipative sites. The retained state includes the field and its coupling history over $[t-\tau, t]$. The readout projects to the spatial mean $m(t) = 64^{-1}\sum_i c_i(t)$, then to the lock-in estimator $\hat{Z} = (2/T_{\mathrm{win}})\int_{\mathrm{window}} m(t)e^{-i\Omega t}\,dt$. Amplitude $\mathcal{A} = |\hat{Z}|$, phase lag $\phi = |\arg \hat{Z}|$, and window-averaged transverse spatial variance V summarize this response. Across 13 delays $\tau \in [0.43, 0.46]$, Table 1 compares amplitude and phase with harmonic-balance targets. Appendix E.4.1 gives sampling settings and separates finite-window estimator bias from the remaining numerical–analytic difference. Spatial concentration uses squared peak amplitude and transverse variance; it does not identify a minimal autonomous state or remove the required history.

Table 1: **Response calibration and separated readout scales.** Consensus errors compare 100-seed means with harmonic targets (maximum over 13 delays); $\mathcal{A}$ is peak amplitude and V is transverse variance. Drive F, noise σ and flock angular noise η use model units. Subscripts P, Q identify the two readouts. Consensus half-values are interpolated crossings of the fraction of trajectories satisfying each criterion through one half; flock half-values are the first sampled noise levels with mean readout below one half. Neither identifies a critical point.

Measurement	**Recorded result**
Consensus response	Relative amplitude error $\leq 2.51 \times 10^{-3}$; absolute phase error $\leq 0.74^\circ$.
Consensus concentration	$\mathcal{A}^2/(\mathcal{A}^2 + V) \geq 0.999$.
Consensus: decreasing drive	$F_Q \approx 0.163$, $F_P \approx 0.118$.
Consensus: increasing noise	$\sigma_Q \approx 0.138$, $\sigma_P \approx 0.310$.
Flock: increasing noise	First sampled values: $\eta_Q = 0.35$, $\eta_P = 0.60$, for 30° alignment on $\eta \in [0.15, 0.70]$.

Separate retrospective drive/noise sweeps at $\tau = 0.43$ hold the original criteria fixed: $P_f = 1$ requires stability, $\mathcal{A} \geq 0.16$, $V \leq 0.05$; $Q_f = 1$ additionally requires $\mathcal{A} \geq 0.22$, $\phi \leq 35^\circ$. Across 26 control values with 50 matched seeds each, 442/1,300 trajectories have $Q_f = 0 < P_f$, with no nesting violations. On the noise axis, all 179 trajectories with $P_f = 1, Q_f = 0$ violate the phase criterion. Nesting guarantees weak ordering; interpolation of the observed fractions quantifies their separation. These sweeps are retrospective. The noise-axis inner crossing has a broad 95% paired-bootstrap interval $[0.075, 0.303]$ (Appendix E.4.1).

Field-driven polar flock. A fixed aligning field and angular-noise loading η act on 256 moving particles; positions and headings evolve by nonlinear local circular-mean coupling (Appendix Eq. E.9). Polarization P_f averages $|256^{-1}\sum_i e^{i\theta_i(t)}|$ over time, then four seeds; Q_f similarly averages the fraction within 30° of the field. The modulus precedes averaging. The readouts are nonnested: numerical order on $\eta \in [0.15, 0.70]$ fails outside it, and Q_f/P_f is not a conditional fraction. Their sampled half-value separation is threshold-dependent, not a boundary trajectory or slow-mode measurement. These averages do not distinguish heading broadening from collective-

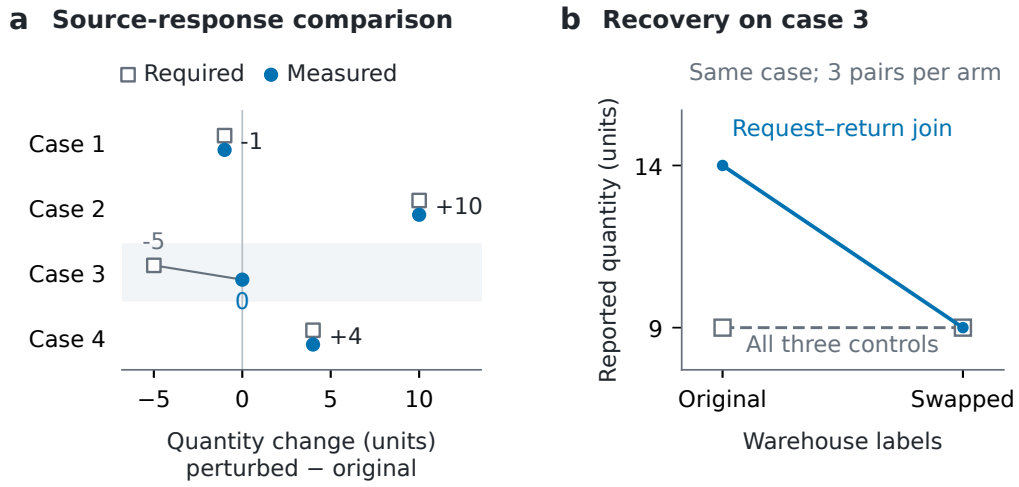


Figure 3: **A missing response and its recovery on the same case. a**, Task-required changes (hollow squares) and measured changes (filled circles) in four constructed inventory cases; case 3 gives zero instead of -5. Small vertical offsets separate coincident marks on categorical rows. **b**, A separate exploratory 24-call test on case 3. Exchanging warehouse labels gives $14/\mathrm{allocate} \rightarrow 9/\mathrm{hold}$ after the request–return join (solid line). Original input, checklist and unrelated-note controls each give $9/\mathrm{hold} \rightarrow 9/\mathrm{hold}$; their coincident outcomes share the dashed line and hollow squares. Each mark represents three identical within-case observations, not independent tasks. Lines connect two input conditions, not times. The join exposes existing associations without adding an answer. Complete outcomes appear in Appendix Tables 15 and 16.

direction fluctuations. Appendix E.4.2 gives a descriptive angular-shape comparison; Appendix Figure 6 gives the full plots.

5 AI RESPONSE MEASUREMENTS AND DIAGNOSTIC INTERVENTIONS

The AI studies link source-producing components, retained communication records and consumers. Graphs of events labeled by type and a delayed-tool request history instantiate these interfaces. The experiments measure responses, test candidate state descriptions and evaluate diagnostic interventions.

Finite contrasts guide an answer-free correction. Four constructed inventory cases use virtual delayed queries with preserved request identifiers. A source contrast exchanges two queries' warehouses in one fixed tool environment, preserving the snapshot and subsequent numerical returns. The reader sees the full request history. In three cases its quantity changes with history at the same snapshot; in case 3 it stays at 9 despite a required change from 14 to 9 (Figure 3a). A rule fixed before model calls selects a request–return join for this missing response, before any repair outcome is observed. The join links returns to requests without discarding records.

Calibration gives 21/24 correct outputs, below the fixed 22/24 criterion; the scheduled validation was not performed. An exploratory follow-up evaluates all four cases in 72 fresh calls. The selected join corrects the faulty case in all three repeats; unchanged retries and a readout checklist, tested alongside the join, correct none. The join makes the request–return associations already in the input explicit, without adding an answer. A separately specified 24-call paired follow-up recovers $14 \rightarrow 9$ in all three join pairs, with both endpoints correct; raw, checklist and unrelated-note pairs each remain $9 \rightarrow 9$. This demonstrates recovery of the task-required finite contrast on the same known case, rather than held-out cases. Ordinary diagnosis selects the same action; fixed joining also succeeds at lower diagnostic call cost (Appendix E.5).

Table 2: **Readout changes in 9 of 12 directions; slopes agree across step sizes in 2 of 12.** Twelve directions span four instances of one template. Graph propagation is imposed; model readouts are measured afresh. Fractions count perturbation directions, not tasks or calls.

Check	Directions	Meaning of the result
Retained-state change	12/12	Euclidean state distance exceeds 0.05 by construction.
Model-readout change	9/12	Support-score change per unit source amplitude exceeds 0.01 in magnitude.
Slope agreement	2/12	Halving the perturbation step changes the slope by at most 25% of the larger magnitude.
Slope sign reversal	4/9	Among the nine above-threshold responses, four reverse sign when the step is halved.

Distances use normalized graph coordinates; the fixed thresholds are not accuracy gains. The three below-threshold slopes occur in one instance and need not be zero. Appendix E.2 gives the exact slope test, call counts and Y/N probability-saturation diagnostics.

Constructed calculation and fixed-record replay. A code-repair graph constructed using endpoint information checks the implemented response composition. A separate deterministic replay of seven model records checks source timing and finite differences; coincident status values yield one numerical response template. These are implementation and replay checks, with complete residuals and controls in Appendix E.2, rather than fresh-model intervention results.

Fresh model readouts after imposed propagation. Four instances of a synthetic three-field template, excluded from calibration, use fresh readouts from MLX 4-bit quantized Qwen2.5-14B-Instruct. A fixed order selects 12 of 15 events; sources and graphs are fixed and recorded before endpoint results are revealed. All four source extractions and 648 endpoint calls are retained, without retries (Table 2). The reader receives numerical supports and a candidate. Separate binary Y/N queries yield within-field normalized Y likelihoods; the reported score averages support for the verifier-selected values across the three fields. This is a support surrogate, not a calibrated joint choice probability. Appendix E.2 gives the prompts, loading and controls.

Propagation imposes the state displacements; the readout alone is measured afresh. Steps of 20% and 10% of baseline source amplitude resolve changes but fail the all-direction linearity criterion. Five directions exceed the endpoint-slope threshold 0.01 and remain sign-stable; only two pass step refinement. This does not isolate whether probe size, nonlinearity or discreteness causes the mismatch. The three-column state-contrast matrix has full column rank by construction, so these contrasts supply no nonzero combination with zero state contrast for a first-order test; the restricted reader interface also imposes access through the support table. Routing controls therefore test the implementation rather than identify mediation or a jointly formed live multi-agent state.

Diagnosis checked against reference calculations. Response diagnosis requires task-specific checks and consumer-input access (Appendix Table 14). Two archived programs compute an integral and an inverse: cumulative flow and its target-volume time, or a calibration curve and its target-response coordinate. Their defects were known before case selection and rule definition. They came from a separate 30-call development experiment with $0/18$ passing eligible accuracy arms. A constructed graph adapter exactly encodes preparation and saved worksheet values and reloads them before consumption; it neither adds information nor corrects arithmetic.

Checks test preparation, inverse range and integral response. A single violation selects the corresponding candidate intervention; multiple violations trigger abstention. Replacement applies a reference formula without modifying the program or calling the model again. The selector receives no reference endpoint or repair outcome; its rules were defined after the archive defects were known.

The archive cases have correct preparation and a defective readout; replacing the other readout or repeating execution corrects neither. Constructed controls retain missed defects, incomplete repair and ambiguous attribution (Table 3); transport, restoration and transplant tests establish adapter input de-

Table 3: **Jointly correct readouts after reference-formula replacement.** Outcome fractions count correct integral and inverse readouts among scored cases, not population repair rates. The constructed and prospective groups include two initially healthy cases each. The unsupported-domain control is unscored. Both interventions use reference formulas: Rule-selected follows the diagnostic rule; Replace both substitutes both formulas.

Case group	**Scored/total**	**Jointly correct/scored**	
		Rule-selected	**Replace both**
Retrospective archived programs	2/2	2/2	2/2
Constructed controls	8/9	4/8	6/8
Prospective four-case pilot	4/4	2/4	4/4

pendence. An archived comparison holds worksheet values and program bytes fixed while changing preparation, which the program also consumes. The differing endpoints confirm this omitted input and reject a worksheet-only state. A separate constructed consumer reads an omitted raw sensor: with the entire graph fixed, its total changes from 1.8 to 2.0. This rejects graph-only sufficiency for that wider input class, and demonstrates how the protocol detects a missing response-relevant variable (Appendix E.2.1). Ordinary diagnostics given the same information can apply these checks.

A prospective pilot fixes four new numeric settings and the unchanged rules before eight model calls. It corrects one integral component but misses both inverse defects because incorrect inverses remain within their allowed ranges. Complete repair is $0/2$ faulty programs; joint correctness stays $2/4$, including the two healthy cases. The fixed positive criterion—full repair and a resolved pre-intervention graph response of the diagnosed readout—fails: the detected integral defect has zero response to the uniform shift. All four cases and control arms remain in Appendix E.2.2.

Exploratory persistence ablations and retrospectively selected spreadsheet paths provide further case-specific comparisons, including a nonmonotone path (Appendices E.2.4 and E.2.3).

6 Discussion and limitations

Paired responses guide forward analysis and inverse tests of retained information (Table 4). The graph study resolves readout changes, but its full-column-rank state-contrast matrix supplies no nonzero state-null combination. More repeats can improve precision; testing adequacy needs new directions or informative matched-state histories. In contrast, equal worksheets with different endpoints reject the worksheet-only representation: an adequate state must distinguish those preparation histories.

The delay counterexample shows why current values omit relevant history; complete history and fixed subsequent inputs give unique continuation under model well-posedness. Separately, consensus calibrates driven responses with history retained throughout, without independently validating an inferred representation.

The delayed-tool join, selected before repair outcomes, restores a two-point contrast on the same known case by rearranging existing associations. It tests presentation, not missing-state discovery. Ordinary diagnosis agrees; fixed joining costs fewer calls. The graph-readout experiment fails its linearity criterion; the prospective program pilot fails its repair criterion (Section 5; Appendix Table 7). In the 89-task development study, primary selection and direct reading have identical correctness; selection costs more. At least one of two specific evidence-reselection responses is available on 35/89 tasks. This fraction concerns those responses; other measurements can remain available. The study does not establish whether missing responses explain the absent decision gain (Appendix E.6).

Scope. The AI studies use accessible, externally specified workflow states. Refined representations need informative new comparisons, fixed consumers and interfaces, and independent error controls. Positive sufficiency, causal mediation, autonomous reduction, cross-model transfer and jointly formed multi-agent states require separate evidence. Intervention-selection benefit requires independent-task comparisons with fixed, response-blind and same-information policies, counting

Table 4: **Response comparisons guide experimental design.** Judgments concern a specified readout, perturbation class and window, under Section 2's matching and measurement-error controls; first-order claims additionally require a qualified local linear approximation. Judgments may coexist across directions. Stochastic readouts require the specified distributional comparison. Bounded support is a possible protocol outcome; the reported examples do not establish an informative positive state-sufficiency result.

Observation	Supported conclusion	Next experiment
Adequate direction coverage; readout-difference bounds, including errors, within the fixed equivalence margin; detectable positive control.	Bounded support on the tested comparisons at the declared resolution.	Extend directions, readouts or windows to test the scope of support.
Exactly equal retained states give a readout difference beyond measurement error.	Reject sufficiency for these histories under the declared conditions.	Separate the violating histories in a revised representation; test it with consumer and interface unchanged.
Approximately matched states violate the independently justified sensitivity-plus-error bound.	Reject the bounded readout class, not the representation alone.	Check the readout class and independent sensitivity bound; tighten state matching.
Neither an informative direction in the state-response kernel nor a matched-state comparison is resolved.	The design leaves adequacy unresolved.	Change source directions or compare different histories at matched states.
Response is detected; slopes fail the step-refinement tolerance.	Finite contrasts; first-order interpretation is unsupported at these steps.	Test finer steps or another readout; recheck resolution.
A task-required contrast is absent despite preserved input records.	Consumer use of those records is a hypothesis to test.	Compare consumer interventions with controls; remeasure both endpoints.

measurement costs; equal information permits equal analytic judgments. Emergent transitions and common empirical response laws require separate tests.

AI use statement. Language models supply source records, evaluations of support for candidate values and programs for the experimental apparatus; they also select evidence and classify claims in the development study (Section 5; Appendix E). AI assisted delayed-tool task design, diagnostic-rule implementation and execution; review of theoretical assumptions and derivations, citation searches, numerical reanalysis, figure preparation, manuscript editing, format conversion, anonymous-export tooling and verification checks. The reported checks reconstruct stored records and analytic examples; AI-generated material is not independent scientific validation. The authors are responsible for the manuscript, claims, evidence and released artifacts.

Reproducibility statement. The appendix supplies assumptions, derivations, protocols and controls. Anonymous code and data reconstruct recorded responses, exact-arithmetic diagnoses, eleven offline cases, the four-case prospective pilot and delayed-tool records (96 calibration/intervention; 24 paired) without a model. Separate anonymous records reconstruct the 89-task development comparisons and request costs. Export documentation identifies record transformations and verification scope. Bootstrap regeneration, broader benchmarks and the complete figure-generation pipeline are excluded (Appendix F).

REFERENCES

Shun-Ichi Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2):77–87, 1977.

Yarden Bakish, Amir Dudai, Roy Ganz, Oren Nuriel, Elad Ben Avraham, Mor Shpigel Nacson, and Ron Litman. Adaptive influence graphs for failure attribution in multi-agent systems. *arXiv preprint arXiv:2608.24361*, 2026. doi: 10.48550/arXiv.2608.24361. URL https://arxiv.org/abs/2608.24361v1. Version 1.

Amine Barrak. Traceability and accountability in role-specialized multi-agent LLM pipelines. *arXiv preprint arXiv:2510.07614*, 2025.

Paul C. Bressloff. Spatiotemporal dynamics of continuum neural fields. *Journal of Physics A: Mathematical and Theoretical*, 45(3):033001, 2012.

Hugues Chaté. Dry aligning dilute active matter. *Annual Review of Condensed Matter Physics*, 11: 189–212, 2020. doi: 10.1146/annurev-conmatphys-031119-050752.

Ning Coeva. Beyond text-passing: Shared cognitive substrates for multi-agent LLM coordination, 2026. URL https://openreview.net/pdf/abde5375768df613a2616a5a644b98b7a97f2b2c.pdf. ICLR 2026 Workshop on MALGAI.

Batu El, Jinhee Paeng, Fatih Dinc, Shiye Su, Mete Erdogan, Aneesh Pappu, Haotian Ye, Wanjia Zhao, Surya Ganguli, and James Zou. Physics of agents: Statistical mechanics predicts collective behavior of AI agents. *arXiv preprint arXiv:2608.16578*, 2026. doi: 10.48550/arXiv.2608.16578. URL https://arxiv.org/abs/2608.16578v2. Version 2, September 9.

Emilia Fridman. *Introduction to Time-Delay Systems: Analysis and Control*. Springer, Cham, 2014.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://jmlr.org/papers/v26/23-0058.html.

J. Gertler. Analytical redundancy methods in fault detection and isolation—survey and synthesis. *IFAC Proceedings Volumes*, 24(6):9–21, 1991. doi: 10.1016/S1474-6670(17)51119-2.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003.

Jack K. Hale and Sjoerd M. Verduyn Lunel. *Introduction to Functional Differential Equations*. Springer, New York, 1993.

Daniel Henry. *Geometric Theory of Semilinear Parabolic Equations*, volume 840 of *Lecture Notes in Mathematics*. Springer, Berlin, 1981. doi: 10.1007/BFb0089647.

Jun Hou, Priya Pitre, Yi Fang, and Xuan Wang. EDGE: Error dependency graph-guided multi-error attribution in multi-agent LLM systems. *arXiv preprint arXiv:2609.01360*, 2026. doi: 10.48550/arXiv.2609.01360. URL https://arxiv.org/abs/2609.01360v1. Version 1.

Ryogo Kubo. Statistical-mechanical theory of irreversible processes. I. general theory and simple applications to magnetic and conduction problems. *Journal of the Physical Society of Japan*, 12 (6):570–586, 1957. doi: 10.1143/JPSJ.12.570.

Ryogo Kubo. The fluctuation-dissipation theorem. *Reports on Progress in Physics*, 29(1):255–284, 1966. doi: 10.1088/0034-4885/29/1/306.

Yoshiki Kuramoto. *Chemical Oscillations, Waves, and Turbulence*. Springer, Berlin, 1984. doi: 10.1007/978-3-642-69689-3.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14, 2001. URL https://proceedings.neurips.cc/paper/2001/file/1e4d36177d71bbb3558e43af9577d70e-Paper.pdf.

Ming Ma, Jue Zhang, Fangkai Yang, Yu Kang, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. DoVer: Intervention-driven auto debugging for LLM multi-agent systems. *arXiv preprint arXiv:2512.06749*, 2025. doi: 10.48550/arXiv.2512.06749. URL https://arxiv.org/abs/2512.06749v3.

M. C. Marchetti, J. F. Joanny, S. Ramaswamy, T. B. Liverpool, J. Prost, Madan Rao, and R. Aditi Simha. Hydrodynamics of soft active matter. *Reviews of Modern Physics*, 85(3):1143–1189, 2013. doi: 10.1103/RevModPhys.85.1143.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for general AI assistants. *arXiv preprint arXiv:2311.12983*, 2023.

Subhadip Mitra. Field-theoretic memory for AI agents: Continuous dynamics for context preservation. *arXiv preprint arXiv:2602.21220*, 2026.

Arthur N. Montanari, Chao Duan, Luis A. Aguirre, and Adilson E. Motter. Functional observability and target state estimation in large-scale networks. *Proceedings of the National Academy of Sciences*, 119(1):e2113750119, 2022a. doi: 10.1073/pnas.2113750119. URL https://doi.org/10.1073/pnas.2113750119.

Arthur N. Montanari, Leandro Freitas, Daniele Proverbio, and Jorge Gonçalves. Functional observability and subspace reconstruction in nonlinear systems. *Physical Review Research*, 4:043195, 2022b. doi: 10.1103/PhysRevResearch.4.043195. URL https://journals.aps.org/prresearch/abstract/10.1103/PhysRevResearch.4.043195.

Hazime Mori. Transport, collective motion, and brownian motion. *Progress of Theoretical Physics*, 33(3):423–455, 1965. doi: 10.1143/PTP.33.423.

Silviu-Iulian Niculescu. *Delay Effects on Stability: A Robust Control Approach*, volume 269 of *Lecture Notes in Control and Information Sciences*. Springer London, London, 2001. doi: 10.1007/1-84628-553-4.

Reza Olfati-Saber and Richard M. Murray. Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on Automatic Control*, 49(9):1520–1533, 2004. doi: 10.1109/TAC.2004.834113.

A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations*, volume 44 of *Applied Mathematical Sciences*. Springer, New York, 1983. doi: 10.1007/978-1-4612-5561-1.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2nd edition, 2009. doi: 10.1017/CBO9780511803161.

Arkady Pikovsky, Michael Rosenblum, and Jürgen Kurths. *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge University Press, Cambridge, 2001. doi: 10.1017/CBO9780511755743.

Sriram Ramaswamy. The mechanics and statistics of active matter. *Annual Review of Condensed Matter Physics*, 1:323–345, 2010. doi: 10.1146/annurev-conmatphys-070909-104101.

Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2017. URL https://auai.org/uai2017/proceedings/papers/11.pdf.

David Ruelle. A review of linear response theory for general differentiable dynamical systems. *Nonlinearity*, 22(4):855–870, 2009.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

Cosma Rohilla Shalizi and James P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104:817–879, 2001. doi: 10.1023/A:1010388907793.

Yiheng Sun, Huifei Wang, Yancheng Zhu, Zhenyu Li, Zebin Zhao, and Yifan Yuan. When "must" becomes "maybe": Constraint weakening in LLM agent workflows. *arXiv preprint arXiv:2608.24569*, 2026. doi: 10.48550/arXiv.2608.24569. URL https://arxiv.org/abs/2608.24569v1. Version 1.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

John Toner and Yuhai Tu. Long-range order in a two-dimensional dynamical XY model: How birds fly together. *Physical Review Letters*, 75(23):4326–4329, 1995. doi: 10.1103/PhysRevLett.75.4326.

Henk J. van Waarde, Jaap Eising, Harry L. Trentelman, and M. Kanat Camlibel. Data informativity: A new perspective on data-driven analysis and control. *IEEE Transactions on Automatic Control*, 65(11):4753–4768, 2020. doi: 10.1109/TAC.2020.2966717.

Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet. Novel type of phase transition in a system of self-driven particles. *Physical Review Letters*, 75(6):1226–1229, 1995. doi: 10.1103/PhysRevLett.75.1226.

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

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Zhaohui Geoffrey Wang. AgentTrace: Causal graph tracing for root cause analysis in deployed multi-agent systems. *arXiv preprint arXiv:2603.14688*, 2026.

Hugh R. Wilson and Jack D. Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical Journal*, 12(1):1–24, 1972.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Zekun Wu, Seonglae Cho, Cristian Enrique Munoz Villalobos, Theo King, Umar Mohammed, Emre Kazim, Maria Perez-Ortiz, Sahan Bulathwela, and Adriano Koshiyama. AgentGraph: Trace-to-graph platform for interactive analysis and robustness testing in agentic AI systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(48):41721–41723, 2026. doi: 10.1609/aaai.v40i48.42393. URL https://ojs.aaai.org/index.php/AAAI/article/view/42393.

Shuyao Xiao, Shengling Wang, Xuan Chen, Ke Chao, Ming Cui, Feifei Qian, Fanlin Meng, Chaoyang Mei, Chaoyong Jiang, Qi Ouyang, and Junxi Yi. When errors become memories: Causal pathway tracing in multi-turn memory-augmented LLMs. *arXiv preprint arXiv:2608.30198*, 2026. doi: 10.48550/arXiv.2608.30198. URL https://arxiv.org/abs/2608.30198v1. Version 1.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, and Qingyun Wu. Which agent causes task failures and when? on automated failure attribution of LLM multi-agent systems. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76583–76599. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25cq.html.

Yue Zhao. GRADE: Graph representation of LLM agent dependency and execution. *arXiv preprint arXiv:2606.22741*, 2026. doi: 10.48550/arXiv.2606.22741. URL https://arxiv.org/abs/2606.22741v2. Version 2, September 12.

Robert Zwanzig. Memory effects in irreversible thermodynamics. *Physical Review*, 124(4):983–992, 1961. doi: 10.1103/PhysRev.124.983.

APPENDIX

The appendix first compares ANM with related approaches, then gives the field and response assumptions, proofs, experimental protocols and complete outcomes. Finite-window response, autonomous reduction, retained-state sufficiency and causal attribution have separate conditions.

A RELATED WORK: DYNAMICS, RETAINED STATE AND INTERVENTION

Workflow, memory and trace foundations. Tool-using and conversational workflows make intermediate records and task outcomes accessible (Yao et al., 2022; Schick et al., 2023; Wu et al., 2023), while benchmarks such as GAIA define endpoint tasks (Mialon et al., 2023). Field-based memory dynamics and shared cognitive substrates address retention and coordination (Mitra, 2026; Coeva, 2026); structured handoffs and execution traces support error analysis (Barrak, 2025; Wang, 2026). These provide architectures and records on which a response protocol can operate. ANM specifies measurements and conditions for assessing a declared state; it does not require a particular memory architecture.

The measurement question. ANM requires source manipulation, candidate-state measurement and separate endpoint evaluation under matched preparation and observation windows (Section 2). Stochastic endpoints require controlled repeated continuations. Prescribed graph dynamics or extracted execution relations can supply a candidate representation; their existence alone does not establish the response relation. Required history and independently consumed records must be represented or retained as explicit alternative paths.

Collective dynamical models. El et al. (2026) measure repeated opinion updates on signed communication networks and fit an Ising-inspired logistic transition law, with persona/question embeddings and interaction coefficients. Their setup refreshes message inboxes each round; their fitted model predicts trajectories and collective regimes from initial opinions and tests transfer to unseen network families. Thus the relevant access is to opinions, network structure and fitting features. Both discrete updates and a continuous-time mean-field extension are developed. The comparison with ANM concerns the target of inference—fitted collective evolution versus testing whether source-to-endpoint responses factor through a candidate retained state—rather than a distinction between discrete and continuous dynamics.

State transformation and cross-turn error propagation. Sun et al. (2026) hold an established workflow constraint fixed, transform its handoff artifact, and restrict an executor to that artifact. They measure preservation separately from forbidden action, including multihop transformations, field-restoration interventions and downstream verification with the artifact held fixed. This distinguishes retained information from whether the endpoint action obeys the constraint. Xiao et al. (2026) model questions, responses and memory across turns, crossing clean/error inputs to the memory updater and next-question generator in four counterfactual trajectories. Access to both mechanisms, memory records and subsequent responses supports estimates of the effects of errors entering through either mechanism and their interaction; probability-level probes additionally use likelihoods. Later histories evolve separately. These are close operational precedents for separating state and endpoint effects; their particular contrasts do not by themselves certify an arbitrary compressed state's sufficiency.

Execution and dependency representations. GRADE (Zhao, 2026) separates execution order from dependencies graded by their observed, declared or inferred origin. Its v2 analysis uses temporal traces, access records and failure labels to test predictive increments over run size, with matched edge-attachment controls. Results depend on the probe and grouping protocol; the within-corpus degree-matched comparison does not establish an attachment advantage. AgentGraph (Wu et al., 2026) links typed entities and relations to trace spans, supports temporal replay, and tests reconstructed relation-level prompts using jailbreak and demographic perturbations. Its causal module attributes measured behavioral changes to components. These methods provide representation, provenance and intervention interfaces relevant to ANM; a graph encoding and a response measured through it carry distinct evidence.

Failure attribution and repair. Who&When (Zhang et al., 2025) evaluates agent/step localization from annotated failure logs using whole-log, sequential and binary-search judgments. AIG (Bakish et al., 2026) constructs adaptive, step-grounded inheritance graphs and lets a reader inspect both graph and original history to localize the decisive error. Both evaluate attribution against annotations. EDGE (Hou et al., 2026) instead learns dependencies between error categories from temporal precedence and probability raising, then validates a subset by repairing source spans and rolling out continuations with tool outputs fixed. Its detector combines validated and selected observational edges; inference must retain that distinction and the conditioning on successful source repair. DoVer (Ma et al., 2025) requires checkpoint restoration and executable message/plan edits, preserves the preceding history, and evaluates resumed runs by success, milestone progress and intervention fulfillment. It explicitly allows multiple useful repairs. These approaches supply substantive diagnostic and intervention evidence. An ANM comparison requires matched information, interventions and scored targets; it does not imply an exclusive diagnostic capability or superiority to these methods.

Activation interventions and mediation. Activation patching measures the effect of replacing an internal activation, commonly with a value from a donor execution (Zhang & Nanda, 2024). Such an interchange intervention is a specific assignment $\mathrm{do}(z = z_{\mathrm{donor}})$; a general admissible clamp $\mathrm{do}(z = v)$ need not use a donor. Attribution patching estimates replacement effects with local gradients, introducing a first-order approximation (Syed et al., 2024). Causal mediation analysis compares specified input and mediator interventions to separate direct and indirect effects (Vig et al., 2020); an isolated patch effect does not supply that decomposition. These approaches can include contextual representations and propagation across token positions and layers. ANM's workflow measurements use accessible communication records and explicitly test a candidate representation over declared sources and observation windows. The comparison concerns access, representation choice, intervention semantics and response resolution, including the histories each design retains. History dependence alone is not a distinction from these methods.

Theoretical precursors and scope. Causal abstraction relates models through mappings of states and permitted interventions, requiring agreement of the corresponding interventional solutions or distributions (Rubenstein et al., 2017; Geiger et al., 2025). That target is richer than agreement for one endpoint on a local source slice; an endpoint factorization alone does not establish a causal abstraction. Structural causal models likewise distinguish observational agreement from responses to interventions on intermediate variables (Pearl, 2009). Functional observability asks whether selected functions of state can be reconstructed from measured outputs and known inputs/dynamics. Its linear row-space condition and nonlinear reconstruction theory provide direct precedents for testing information lost under a projection (Montanari et al., 2022a;b). ANM's condition $\ker \mathcal{K}_{z,T} \subseteq \ker \mathcal{K}_{O,T}$ uses the same linear factorization principle on source-response maps. It remains local and is automatic when the candidate-state response has full column rank. Stochastic bisimulation preserves rewards and block-transition probabilities for every action (Givan et al., 2003); predictive state representations summarize action–observation history through predictions of future tests, with recursive updates (Littman et al., 2001). Causal states group histories with the same conditional distribution of futures and give minimal predictive representations (Shalizi & Crutchfield, 2001). ANM instead tests a declared coordinate for specified sources, readouts and finite windows; it does not establish such predictive minimality. These precedents clarify why retaining a useful coordinate is weaker than establishing the relevant equivalence over future behavior. These connections locate ANM's contribution in its specified response measurements and tests, rather than a new general principle of abstraction.

Data informativity asks whether every system consistent with the observed data has a target property, or admits a common controller; some objectives require less information than unique system identification (van Waarde et al., 2020). This provides a direct precedent for asking what an experiment can decide. ANM applies the question to separately measured state and endpoint responses, with explicit tests for unresolved contrasts and uninformative perturbation designs, rather than introducing a general theory of data sufficiency.

B PHYSICAL SYSTEM AND FIELD LAW

B.1 SCOPE AND ORGANIZATION

Following the related-work comparison (Appendix A), this section defines the delayed state, conditional autonomous response and representation tests. Sections C and D develop the optional nested-readout case and distinguish finite-window response from stationary susceptibility. Section E records the experiments and their controls; Section G collects notation.

The construction is shared across realizations; state coordinates, response functions and dynamics remain system-specific. Retained-state adequacy and autonomous closure are separate conditions. Each experiment identifies whether its evidence is a constructed calculation, fixed-record replay, fresh readout measurement or steady-state sweep.

Throughout, "admissible" means that the instance class, interface, loading coordinate, stopping rule and readout definitions remain fixed during the comparison. Standard symbols are used for spaces and maps; Im and $\ker$ denote image and kernel. Hats on response quantities denote measured or finite approximations. Additional conventions are listed in Section G.

B.2 FORMAL DELAYED FIELD-COUPLED MODEL

Retained information changes through local interactions, relaxation and delayed inputs. Separating these contributions fixes where a source perturbation enters and which state must be followed before the endpoint is read. Let $H = L^2(X, \mathbb{R}^q)$ or another separable Hilbert field space over a domain X of physical positions or declared graph sites, with q realization-specific channels. These may represent physical activity or, in the AI examples, weights assigned to candidates, the strength of conflicting evidence, inhibition or tool availability. Let $c_t \in H$ denote the shared coordination field, and let each local operator i carry an internal memory state $m_i(t) \in M_i$, where M_i is a finite-dimensional or Banach memory state space on the local neighborhood used in this section. Its net contribution to the shared field is a composite source $J_i(c_t, m_i(t)) \in H$.

To track how local writes and delayed inputs alter the shared field, we model its evolution in Active Neural Matter (ANM) as the sum of intrinsic medium evolution, local sources and externally supplied delayed signals:

$$\partial_t c_t = Lc_t + \sum_{i=1}^{N} J_i(c_t, m_i(t)) + \sum_{\ell=1}^{N_\tau} E_\ell y_\ell(t - \tau_\ell), \tag{B.1}$$

where L is the autonomous evolution generator of the medium, J_i is the composite source generated by the ith modeled local operator, N_τ is the number of delayed channels represented outside the modeled local operators, $y_\ell(t - \tau_\ell)$ is a delayed record in a signal space Y_ℓ, and $E_\ell : Y_\ell \to H$ is a bounded delayed-input coupling that maps this record into a field-space increment. The delayed signals in the last term have generating processes outside the modeled local operators, such as verifier corrections, tool outputs, environmental records, or externally applied delayed drives.

When the mechanism producing the source is resolved, a field-coupled local operator is specified by a read operator $R_i : H \times M_i \to O_i$, a local update map $\pi_i : O_i \to U_i$, a write operator $W_i : U_i \to Y_i^{\mathrm{src}}$, a source-coupling operator $G_i : Y_i^{\mathrm{src}} \to H$, and a memory-update rule Ψ_i. In that resolved description,

$$J_i(c, m_i) = G_i W_i(\pi_i(R_i(c, m_i))) .$$

Here O_i is the local readout space, U_i the update-output space, and Y_i^{src} the source space exposed by operator i.

The two source sums distinguish modeled generators from externally supplied signals. Internal feedback, including delayed feedback represented in memory or augmented histories, belongs to J_i. A delayed tool, verifier or environmental record belongs to $E_\ell y_\ell$ only when its generator is outside the modeled operators. Accordingly, delayed consensus carries its feedback history internally and requires no additional external delayed-input channel for that coupling.

The terms in Eq. (B.1) have the following physical roles. The operator L is the generator of the autonomous medium dynamics: it describes how field already present in the medium spreads, relaxes,

decays, or mixes before any new operator source or delayed source is added. In a continuous spatial realization one may have, for example, $L = D\Delta - \Gamma$, with diffusion tensor D and dissipative rate matrix Γ; on a finite graph one may have $L = -DL_G - \Gamma$, where L_G is the graph Laplacian. In the abstract Hilbert setting used in this section, the corresponding assumption is that L generates a strongly continuous contraction semigroup on H.

The source-coupling map G_i places the local write $W_i(\pi_i(R_i(c_t, m_i(t))))$ into the field. For a PDE it may be an injection kernel,

$$(G_i w)(x) = \int_X g_i(x, \xi) w(\xi)\, d\xi,$$

where w denotes the written source density in this formula. On a finite graph, G_i selects the sites and channels receiving the write.

This source-injection kernel is distinct from the Green function of the field equation. The Green function, or equivalently the semigroup generated by L, describes how an injected source subsequently propagates and relaxes in the medium. In a linearized PDE realization the contribution of an operator source has the form

$$\delta c(t, x) = \int_0^t \int_X K_L(t - s; x, x')\, (G_i \delta w_s)(x')\, dx'\, ds,$$

where K_L is the Green kernel of $\partial_t - L$. Here K_L propagates a specified source contribution through the autonomous medium. The response of the coupled system to an independent perturbation also includes the derivatives of the operator sources, memory dynamics and internal delayed feedback; these enter the augmented linearization below. Thus g_i specifies how a local write enters the field, whereas K_L specifies how the field carries that injected perturbation forward in time.

The map E_ℓ couples a delayed signal generated outside the modeled local operators into the same field. It maps a verifier output, tool state, environmental observation, externally applied drive, or other delayed record $y_\ell(t - \tau_\ell)$ into a field forcing term. The boundedness condition means that there is a finite gain

$$\|E_\ell y\|_H \le \beta_\ell \|y\|$$

on the relevant signal space. For an input bounded by M_ℓ on the sampled history interval, the forcing magnitude is at most $\beta_\ell M_\ell$. Its squared norm integrates to at most $T\beta_\ell^2 M_\ell^2$ over a window of length T. Bounded coupling alone neither bounds an arbitrary input nor establishes long-time boundedness of the driven state. Table 5 summarizes the physical interpretation of the field-equation terms L, J_i and E_ℓ.

Table 6 gives concrete realizations.

The closed product-space system used for well-posedness also specifies the memory evolution. In continuous time one may write

$$\dot{m}_i(t) = \Psi_i(c_t, m_i(t), \eta_i(t)), \qquad i = 1, \dots, N, \tag{B.2}$$

where $\eta_i(t)$ denotes bounded apparatus input available to the ith local operator. The continuous-time well-posedness result below applies to memory laws satisfying its stated hypotheses. A discrete or hybrid memory update requires a separately specified recurrence or jump law; an unspecified interpolation does not make it a solution of Eq. (B.2). If a memory trajectory is externally recorded rather than dynamically updated, it is treated as a bounded prescribed input, and well-posedness of the remaining dynamical components requires the corresponding source regularity conditions.

Equivalent source terms. Fix the same initial field, the same external inputs and all other field-source histories. If two local mechanisms expose identical effective source histories,

$$J_i(c_t, m_i(t)) = J_j(c_t, m_j(t))$$

for almost every $t \in [0, T]$, variation of constants gives the same field trajectory on that horizon. Every declared readout depending only on this field trajectory therefore agrees. For readouts that also use separately retained memories or other coordinates, equal additional inputs are sufficient for agreement; otherwise the endpoints must be compared through the full readout map.

Table 5: **Physical roles of the field-equation terms.** Equation (B.1) distinguishes autonomous medium dynamics L, composite local sources $J_i = G_i W_i(\pi_i(R_i))$ when their mechanisms are specified, and externally generated delayed inputs $E_\ell y_\ell$. In the example generators, D is the diffusion coefficient, Γ the relaxation operator, and L_G the graph Laplacian.

Symbol	Physical meaning	Typical realization
L	autonomous propagation, relaxation, and dissipation of the shared medium	diffusion–decay operator $D\Delta - \Gamma$, graph Laplacian generator $-DL_G - \Gamma$, or Markov/contraction generator
J_i	composite source generated by modeled local operator i	feedback force, actuator drive, graph-node source, or typed evidence increment
R_i	local readout available to operator i from the field and memory	projection, local averaging, attention window, tool-state readout
π_i	local update or feedback law used by operator i	controller law, consensus update, model inference, decoding, or tool choice
W_i	local write produced by operator i's update rule	signed source density, graph-node injection vector, channel-specific evidence update
G_i	source-coupling operator placing the local write into the medium	spatial source kernel $g_i(x, \xi)$ or finite source matrix on a graph
E_ℓ	finite-gain coupling from delayed input outside the modeled local operators into the medium	verifier, tool, environment, or measurement channel with delay τ_ℓ

Table 6: **Realization-specific source mechanisms.** The composition $J_i = G_i W_i(\pi_i(R_i))$ in Eq. (B.1) separates reading, updating, writing and injection; the listed writes precede injection by G_i. Consensus delay belongs to the internal history; delayed tool or verifier records enter through E_ℓ only when their generators are external to the modeled operators, even if an operator action caused the record.

Realization	R_i (read)	π_i (update)	W_i (write)
Driven delayed consensus network (Appendix E.4.1)	neighbor-state read	consensus update rule	delayed network broadcast
Field-driven polar flock	neighbor-heading read	alignment-to-field update	advected position and heading
Distributed AI workflow, language-model operator	family-specific state and trace readout	family-specific state update under the fixed protocol	write into the local operator source space Y_i^{src}, then injection through G_i

Independently specified read, update, write and injection maps provide selective intervention targets. Equal exposed sources do not identify these mechanisms; component diagnosis requires additional measurements or selective interventions. Retriever calls, verifier checks and tool outputs instantiate source or input channels in AI workflows.

Distributed-system apparatus and experimental instance. A distributed-system apparatus is the experimental setup providing external memory, tool state, evaluation loops, and execution control around one or more local operators. AI systems are one concrete example. An experimental instance a is one fixed task or experiment presented through such an apparatus under fixed interface, permissions, and stopping semantics. This label specifies the preparation, not a dynamical state variable of the coordination law; reduced coordinates are defined separately.

Shared coordination state. The shared coordination state is a retained coarse-grained state description $z_t = G_f^{\mathrm{cg}}(s_t)$ that carries information from earlier states or coupling signals across local interactions. Here s_t includes the field, memories and required histories. In the AI examples, its channels may carry unresolved task state, conflicting evidence, delayed evidence and pending corrective actions. In measured neighborhoods where the source-to-endpoint response factors through this coordinate, it is the retained state used by the boundary theory. The reduced order parameters are defined only after this stage.

Preparation, coarse-graining, and endpoint readout maps. For a fixed family f, let a denote a raw task or experimental instance. Preparation $P_f^{\mathrm{prep}}(a) = s_0$ specifies the initial field, memories and histories; subsequent dynamics and inputs determine $s_T \in X_f$. The map $G_f^{\mathrm{cg}} : X_f \to Z_f$ gives the candidate retained coordinate $z_T = G_f^{\mathrm{cg}}(s_T)$. The endpoint readout is $O_T(s_T)$. Where it factors through the candidate, write $O_T = \bar{O}_T \circ G_f^{\mathrm{cg}}$, with $\bar{O}_T$ defined on retained coordinates. This factorization is a condition to test. A nested pair (P_f, Q_f), when declared on z, gives separate operating criteria; its reduced variables (w, r) are defined where the ratio exists. Preparation selects the initial condition rather than defining order parameters.

Window observations as terminal readouts. A time average or lock-in estimate reads a trajectory, not just its last field value. To write it as $O_T(s_T)$, include the required record in the full state. Any source or other input read directly by the endpoint must also be included in s_T; it need not belong to the candidate z_T. For an integral measurement, append an observation accumulator $\mathbf{a}_t$ satisfying $\dot{\mathbf{a}}_t = \psi(c_t, t)$, $\mathbf{a}_0 = 0$; then $\mathbf{a}_T = \int_0^T \psi(c_t, t)\,\mathrm{d}t$. The specified weighting function ψ can select the measurement window and the real and imaginary lock-in components. Amplitude and phase are functions of these components; a differentiable phase uses a continuous local branch at nonzero amplitude. More general path readouts require their corresponding trajectory record. This observation augmentation records the dynamics without feeding back into them. It neither certifies a candidate z_T nor asserts autonomous stable dynamics for the observation coordinates.

B.2.1 History, projection and conditional autonomous response

The finite-window response follows perturbations through the full state. An autonomous equation for fewer coordinates needs a separate closure assumption. The distinction is particularly important when past signals affect future motion.

Include internal history before projecting. For each internally generated delayed quantity q_a, retain $\xi_a^t(\theta) = q_a(t + \theta)$ on $[-\tau_a, 0]$. Together with the field and local memories, these histories form

$$s(t) = (c_t, m_1(t), \ldots, m_N(t), \xi_1^t, \ldots, \xi_{N_{\mathrm{int}}}^t) \in X_f.$$

Each q_a must be a specified causal function of modeled variables; any additional signal generator needs its own state equation. Internal histories evolve by $\partial_t \xi_a^t = \partial_\theta \xi_a^t$, with boundary value $\xi_a^t(0) = q_a(s(t))$. Field evolution, memory laws $\dot{m}_i = \Psi_i(c, m_i, \xi; u)$, and these history shifts define

$$\dot{s} = F_{\mathrm{full}}(s; u, y^t). \tag{B.3}$$

The histories y^t of external signals remain prescribed inputs. Section B.3 specifies the continuous-history space for mild solutions; the displayed derivatives additionally require classical solutions in

the generator domain. Observation records are appended when required by the readout (Section B.2). Internal consensus delay therefore remains in s even without external delayed inputs.

Declare a candidate $z = G_f^{\mathrm{cg}}(s) \in Z_f$ from its memory and coupling roles before testing source-to-readout factorization. The current field alone can omit histories or memories. Neither the projection nor a successful endpoint factorization establishes autonomous dynamics for z.

Closure on an invariant branch. Fix control u and stationary external input $\bar{y}$. Suppose a locally invariant branch admits a C^1 parametrization $s = h_u(z)$ with $G_f^{\mathrm{cg}} \circ h_u = \mathrm{id}$, where G_f^{cg} is also C^1. Branch trajectories must be classical solutions: fields lie in $D(L)$, and histories have compatible derivatives and boundary traces. Assume $F_{\mathrm{full}}(h_u(z); u, \bar{y})$ is tangent to this branch. Then $\mathrm{D}_s G_f^{\mathrm{cg}}\, \mathrm{D}_z h_u = I$, and the chain rule gives

$$\dot{z} = \mathrm{D}_s G_f^{\mathrm{cg}}(h_u(z)) F_{\mathrm{full}}(h_u(z); u, \bar{y}) =: f_u(z). \tag{B.4}$$

This is closure on the assumed branch. Attraction to the branch, needed to approximate nearby off-branch trajectories, is an additional dynamical condition. Fixed-parameter invariance does not establish tracking of $h_{u(t)}$; arbitrary external forcing requires a controlled description or inclusion of its generator in s. Finite-window measurements require neither this reduction nor a potential representation.

Stable equilibrium and discrete response. For a finite-dimensional closed drift that is C^1 in state and control, let z_* be a stationary point at u_*. Linearization gives

$$\delta\dot{z} = A_z \delta z + B_z \delta u, \qquad A_z = \mathrm{D}_z f(z_*, u_*), \quad B_z = \mathrm{D}_u f(z_*, u_*), \tag{B.5}$$

where $f(z, u) = f_u(z)$. If A_z is Hurwitz, the implicit-function theorem gives a local stationary branch, and differentiating $f(z_*(u), u) = 0$ yields

$$\delta z_* = \chi_* \delta u, \qquad \chi_* = -A_z^{-1} B_z. \tag{B.6}$$

Invertibility suffices for the local branch; stability also lets a persistent small source approach this response. Tracking a moving branch adds $-\chi_*(u)\dot{u}$ to the deviation equation and needs a separate bound.

For a verified closed retained-state transition, with a differentiably factored readout $\bar{O}_T$, write

$$\begin{aligned} \delta z_t &= A_t^{\mathrm{ret}} \delta z_{t-1} + B_t^{\mathrm{ret}} \delta u_t, \\ \Phi_{\mathrm{ret}}(T, t) &= A_T^{\mathrm{ret}} \cdots A_{t+1}^{\mathrm{ret}}, \\ \delta O_T &= \mathrm{D}_z \bar{O}_T \left[\Phi_{\mathrm{ret}}(T, 0) \delta z_0 + \sum_{t=1}^{T} \Phi_{\mathrm{ret}}(T, t) B_t^{\mathrm{ret}} \delta u_t \right]. \end{aligned} \tag{B.7}$$

Here $\Phi_{\mathrm{ret}}(T, T) = I$; changing initial history can make $\delta z_0 \neq 0$. For time-independent A, B with $\rho(A) < 1$, a persistent tangent source has the time-summed response

$$\chi_*^G = \sum_{k=0}^{\infty} A^k B = (I - A)^{-1} B.$$

Under exact zero-order-hold discretization, $A = e^{A_z \Delta t}$ and $B = \int_0^{\Delta t} e^{A_z s} B_z \, \mathrm{d}s$, so $(I - A)^{-1} B = -A_z^{-1} B_z$. This follows from $\int_0^{\Delta t} e^{A_z s} \, \mathrm{d}s = A_z^{-1}(A - I)$. An approximate propagator requires a consistent source discretization. The stable endpoint response is therefore

$$\delta O = (\mathrm{D}_z \bar{O}) \chi_* \delta u \quad \text{or} \quad \delta O = (\mathrm{D}_z \bar{O}) \chi_*^G \delta u. \tag{B.8}$$

The constructed code-repair graph and fixed-record replay check this composition under prescribed graph dynamics; they do not identify an autonomous state of a language model.

B.2.2 Minimal representation of a terminal response

For fixed sources, a fixed terminal readout and a finite horizon, the linear response identifies a minimal representation of the reached terminal displacements visible to that readout. This construction

differs from independently declaring a physical retained state. For $\delta z_t = A\,\delta z_{t-1} + B\,\delta u_t$ with zero initial displacement, write $C_T = \mathrm{D}_z \bar{O}_T$ and define

$$\mathcal{R}_T = \mathrm{Im}[B, AB, \ldots, A^{T-1}B], \qquad \mathcal{N}_T = \ker C_T.$$

For time-dependent transitions, $\mathcal{R}_T$ is instead the image of the finite-window source-to-state map. The quotient $\mathcal{R}_T/(\mathcal{R}_T \cap \mathcal{N}_T)$ is isomorphic to $C_T\mathcal{R}_T$: it retains exactly the terminal displacements that this readout distinguishes. Displacements differing only by an invisible component are identified. That component may matter for another readout or later time. The quotient need not carry an autonomous transition. A dynamical minimal realization instead requires reachability and unobservability for a specified sequence of readouts, with the corresponding invariant subspaces. In the driven consensus network the spatial response is carried by the drive-locked consensus mode, which the spatial-mean readout measures; the delay line remains part of the retarded state. In the constructed code-repair graph, the 32-coordinate graph state is retained as an implementation representation. Identity observation under a constructed full source basis does not establish an empirical minimal realization for the declared endpoint pair $O = (P_f, Q_f)$. An effective dimension may be summarized from a specified nonzero response matrix by a specified rank functional. For example, the participation ratio is

$$r_{\mathrm{PR}} = \frac{\left(\sum_k \sigma_k^2\right)^2}{\sum_k \sigma_k^4}$$

of the modal energies σ_k^2 (the squared singular values) of $\chi_*^G = (I - A)^{-1}B$, or of a finite-window map at a stated horizon. The retained-state and endpoint-projected maps generally have different ranks: projection by $\mathrm{D}_z\bar{O}_T$ can lower algebraic rank, and even an injective but non-isometric readout reshapes the singular-value spectrum. A minimal representation of the terminal response is defined relative to its declared source and terminal readout; identity observation of the retained state alone does not establish that representation for a different endpoint.

Worked example: driven delayed consensus. For the network of Eq. (E.1) the declared source is the spatially uniform drive and the declared readout is the lock-in of the spatial mean $m = N^{-1}\sum_i c_i$. The normalized ring coupling W is circulant with unit row sum, so it fixes the consensus (zero) Fourier mode and maps the zero-sum transverse subspace into itself. The uniform source therefore reaches only the consensus mode and the mean readout observes only the consensus mode (the transverse source response and mean-readout sensitivity are zero to numerical precision). Hence the reachable quotient has only the consensus spatial mode; the $N - 1$ transverse spatial modes carry no source-to-readout response. The delay history remains part of the retarded state. The deterministic driven spatial response is rank one. With transverse noise, the amplitude/variance diagnostic is at least 0.999. Under the weights defined in "Reduction accuracy" (Section E.4.1), the participation bounds at the least-concentrated delay both round to 1.002; this is a bounded diagnostic from aggregate weights, not a measured spectrum or an exact rank. Inside the surviving mode the harmonic-balance susceptibility $\chi(\Omega, \tau) = (a - Ke^{-i\Omega\tau} + i\Omega)^{-1}$ fixes the steady harmonic-response amplitude and phase, with analytic amplitude $\mathcal{A}_{\mathrm{th}} = F|\chi(\Omega, \tau)|$. The measured amplitude, phase and transverse variance form the finite-window summary $\boldsymbol{y}_{\mathrm{cons}} = (\mathcal{A}, \phi, V)$; they are not an autonomous state derived from the terminal-response quotient. That quotient is relative to the declared source map and terminal readout. The next definition specifies how evidence histories are encoded as sources in an experimental realization.

Evidence-to-field source map. In finite experimental realizations, the source-construction map ends in a numerical source term for the field. For an observed evidence history $\eta \in E_f^{\le t}$, the source construction is written as the sequence

$$\eta \xrightarrow{\xi_f} e_f^{\mathrm{prop}} \xrightarrow{A_f^{\mathrm{prep}}} e_f^{\mathrm{prep}} \xrightarrow{A_f^{\mathrm{adm}}} e_f^{\mathrm{adm}} \xrightarrow{J_f} j_f(\eta) \in Y_f^{\mathrm{src}}.$$

Here Y_f^{src} is the family-specific source-injection space; ξ_f extracts proposed source events from observed evidence; A_f^{prep} calibrates those events at the schema level, preserving visible pre-verifier evidence, bounding additional support amplitudes by visible evidence when such a bound is specified before evaluation, and excluding endpoint labels from the source term; A_f^{adm} admits the source events that are allowed to act on the field; and J_f converts the admitted event into field source amplitudes, sites, channels, and operator tags. Equivalently, the total source map is

$$j_f(\eta) := J_f\Big(A_f^{\mathrm{adm}}\Big(A_f^{\mathrm{prep}}(\xi_f(\eta))\Big)\Big) \in Y_f^{\mathrm{src}}.$$

Family dependence enters through the evidence process, action set, source-event alphabet, allowed relation types, and verifier; the route from preparation to field is common across deterministic simulations, recorded-data interfaces, and language-model interfaces.

B.2.3 When the source-event representation is sufficient

A sufficient source record must preserve every distinction that an admissible continuation can expose at the readout. The source-event chain ends in the numerical field source $j_f(\eta)$. After admission, source events become source amplitudes, sites, channels, and operators in the field space that drives the reduced dynamics, distinct from endpoint labels.

Source-representation sufficiency. Fix an admissible neighborhood, the observable pair (P_f, Q_f), and the endpoint readout $O_T(s_T)$ at a fixed horizon T. Let $\rho_f : E_f^{\le t} \to R_f^{\mathrm{rep}}$ be a finite representation of evidence histories. Define $\eta_1 \equiv_f \eta_2$ when replacing $j_f(\eta_1)$ by $j_f(\eta_2)$ in any admissible continuation inside the neighborhood leaves all verifier-, readout-, and observable-relevant quantities unchanged: the induced values of P_f, Q_f, and the endpoint readout O_T coincide within the declared calibration. The reduced coordinates are evaluated from the common calibrated pair: $w = P_f$, and $r = Q_f/P_f$ only where the common P_f is positive. Here "coincide" means equality after applying the same fixed calibration or quantization map, so $\equiv_f$ is an equivalence relation rather than a pairwise tolerance test. Then ρ_f is sufficient for the local source representation if and only if no two inequivalent histories are assigned the same finite representation,

$$\rho_f(\eta_1) = \rho_f(\eta_2) \quad \Longrightarrow \quad \eta_1 \equiv_f \eta_2.$$

The coarsest sufficient representation is obtained when each representation class is exactly one equivalence class of $\equiv_f$.

Derivation. If two inequivalent evidence histories are assigned the same finite representation, then any source construction that factors through the finite representation assigns them the same numerical source term although some admissible continuation separates them through the verifier, the endpoint readout, or the nested observables. Such a representation cannot reproduce all specified continuation readouts. Conversely, if each representation class contains only histories equivalent under $\equiv_f$, then, for each fixed admissible continuation, its calibrated observable and endpoint values are functions of the representation. This establishes readout sufficiency on those tests; autonomous evolution of the representation requires a separate closure condition.

A source map retaining only the scalar count of failed verifier calls illustrates how an insufficient representation can arise. It identifies a history in which the verifier refutes the active relation supporting the selected candidate with a history in which a delayed missing-evidence warning concerns an inactive candidate. The two histories can carry the same scalar count while coupling to different sites, signs, or delay channels in the source space Y_f^{src}. The first may suppress Q_f or alter the endpoint readout O_T; the second may leave the local (P_f, Q_f) branch unchanged or perturb a different recovery channel. The scalar representation is insufficient if an admissible continuation distinguishes these histories through P_f, Q_f or O_T. Different source locations or directions alone do not establish that distinction.

The field-source representation Y_f^{src} is refined only when an admissible continuation distinguishes histories that it merges. Distinct source locations are required only when inequivalent histories remain separated by their location of action on the retained state. Distinct channels are required only when histories acting at the same retained location remain separated by their response direction, for example support, contradiction, delayed evidence, inhibition, or a pending corrective action. Additional operator structure is required only when histories with the same retained location and response direction still propagate differently under the reduced dynamics. Equivalently, a collapsed coordinate is admissible only when all source differences inside its representation class are invisible both to the observable response and to the endpoint readout map on the neighborhood.

B.2.4 A falsifiable retained-state criterion

The preceding criterion concerns preparation of the source representation. A separate question is whether a declared retained state contains every source-induced distinction that can affect the endpoint. This question has a direct local test. Let $\mathcal{U}_T$ be the finite-dimensional measured perturbation

space over the observation window, and let $\mathcal{Z}_T$ and $\mathcal{Y}_T$ denote the corresponding vector spaces of retained-state and endpoint variations. Let

$$\mathcal{K}_{z,T} : \mathcal{U}_T \to \mathcal{Z}_T, \qquad \mathcal{K}_{O,T} : \mathcal{U}_T \to \mathcal{Y}_T$$

be the first-order source-to-state and source-to-endpoint maps at the declared operating point.

For the full-state recursion, repeated substitution gives

$$\delta s_T = \Phi(T,0)\delta s_0 + \sum_{t=1}^{T} \Phi(T,t) B_t \delta u_t.$$

Thus each perturbation is propagated from its injection time to the endpoint, and an initial-history change is propagated from time zero. Apply $\mathrm{D}_s G_f^{\mathrm{cg}}$ or $\mathrm{D}_s O_T$ to obtain the two response maps, without assuming that the second factors through the first. Preparation variations, when allowed, belong to the declared perturbation space along with subsequent source changes. The zero-initial-displacement formula in the main text is the fixed-preparation case.

First-order retained-state criterion. There is a unique linear readout map $\mathcal{H}_T^{(1)} : \operatorname{Im} \mathcal{K}_{z,T} \to \mathcal{Y}_T$ satisfying

$$\mathcal{K}_{O,T} = \mathcal{H}_T^{(1)} \mathcal{K}_{z,T}$$

on the measured perturbation space if and only if

$$\ker \mathcal{K}_{z,T} \subseteq \ker \mathcal{K}_{O,T}.$$

Uniqueness is only on the state displacements reached by the measured perturbations; the result makes no claim about unprobed state directions.

Derivation. If the factorization holds, every perturbation that leaves the retained state unchanged also leaves the endpoint unchanged. Conversely, under the kernel inclusion define $\mathcal{H}_T^{(1)}(\mathcal{K}_{z,T}\delta u) := \mathcal{K}_{O,T}\delta u$. Two perturbations producing the same state displacement differ by a vector in $\ker \mathcal{K}_{z,T}$; the inclusion makes their endpoint displacements equal. The resulting map is therefore well-defined and linear, and fixes $\mathcal{H}_T^{(1)}$ uniquely on $\operatorname{Im} \mathcal{K}_{z,T}$.

Delayed dynamics give a physical example in which this first-order criterion rejects an insufficient retained state. For $\dot{x}(t) = -ax(t) - kx(t-\tau) + u(t)$, with $k \neq 0$ and $\tau > 0$, suppose the retained coordinate is truncated to the present value $z(t) = x(t)$ and the endpoint reads $x(t+\Delta)$ for a sufficiently small $\Delta > 0$. If the declared perturbations admit a history variation with $\delta x(t) = 0$ and $\delta x(t-\tau) \neq 0$, holding the subsequent input fixed gives $\delta \dot{x}(t) = -k\delta x(t-\tau) \neq 0$. The nearby endpoint therefore changes along a reached null direction of the present-value coordinate. Equal current values are not equal initial histories for a delay equation, so this does not contradict uniqueness of its solutions. On this perturbation class, histories with the same permitted later inputs can require different endpoints. A sufficient retained representation must preserve information distinguishing those histories; this does not require a unique state representation or retention of a particular raw history record.

The criterion remains local and first order, rather than an exact nonlinear factorization theorem. Let $\mathsf{z}_T(u)$ and $\mathsf{o}_T(u)$ denote the nonlinear retained-state and endpoint maps from a source history u. Exact sufficiency on a finite neighborhood instead requires that any two source histories producing the same retained state also produce the same endpoint:

$$\mathsf{z}_T(u_1) = \mathsf{z}_T(u_2) \quad \Longrightarrow \quad \mathsf{o}_T(u_1) = \mathsf{o}_T(u_2).$$

For C^1 maps, constant rank of z_T and kernel inclusion throughout a neighborhood imply factorization on a sufficiently small local neighborhood. To see why, use constant-rank coordinates (x, y) in which z_T depends only on x and distinguishes nearby x values. The y directions lie in $\ker D\mathsf{z}_T$, so kernel inclusion gives $D_y \mathsf{o}_T = 0$. On a sufficiently small product neighborhood, o_T is independent of y and is a function of z_T. Without the constant-rank condition, pointwise kernel inclusion alone is insufficient: $\mathsf{z}(u) = u^2$ and $\mathsf{o}(u) = u^3$ satisfy the derivative-kernel condition at every scalar u, yet the arbitrarily close inputs $u = \epsilon$ and $u = -\epsilon$ have the same retained state and different endpoints.

From measured response to diagnosis. The criterion concerns a declared state, not a causal fault label. Fix the source intervention, state units, observation window and endpoint rule before comparison. Estimate the state and endpoint contrasts separately, together with their measurement resolution. Resolved changes in both establish a response association under that source intervention; they do not establish that the measured state mediates it. A resolved state change with no resolved endpoint change can reflect cancellation or robustness. If neither response is resolved, the measurements provide no localization at that precision. Testing an unchanged state requires an informative null direction or matched-history contrast and an endpoint positive control. Approximate matches require a justified error allowance; stochastic readouts require repeated controlled continuations and prespecified equivalence bounds. Finitely many tests with no detected violation support only the tested inputs; they do not establish global sufficiency.

For an approximate match with a deterministic readout, suppose the true state difference is bounded by ε_z, the error in the measured endpoint contrast $\widehat{\Delta O}$ by ε_O, and candidate readouts $O = h(z)$ have a Lipschitz bound L_h on the compared states justified independently of the tested endpoint contrasts. Factorization then requires $\|\widehat{\Delta O}\| \le L_h\varepsilon_z + \varepsilon_O$. Exceeding this bound rejects that candidate class. Without a justified sensitivity bound or other continuity control, an approximate state match alone cannot reject every possible readout of that state. An exact match sets $\varepsilon_z = 0$; stochastic endpoint distributions need their own specified comparison and uncertainty control. For example, $z = u$, $O = 2u$ preserves the complete input, yet any distinct pair rejects the class $L_h \le 1$ when $\varepsilon_z = |\Delta u|$ and $\varepsilon_O = 0$. The state needs no extra coordinate: the readout is more sensitive than the tested class allows. Such a rejection calls for checking the readout assumptions and state matching before refining the representation; the same observed violation cannot independently justify a replacement bound.

Inverse use: distinctions a representation must preserve. Fix a family of continuation experiments admissible for every compared history, including sources, readouts, windows and noise laws. Two histories are *response-equivalent* when their specified response quantities agree in every experiment. For a distributional claim, agreement concerns conditional readout laws; equal means suffice only for a mean-response claim. Include baseline readouts when endpoint levels, rather than contrasts alone, are the target. Exact equality defines an equivalence relation, whose classes form the coarsest partition preserving these tests. A sufficient representation may split a class but cannot merge distinct classes: a resolved same-state violation therefore identifies histories it must separate. This is a constraint on refinement, not a reconstruction algorithm or a finite-dimensional autonomous realization. Causal states and predictive-state representations provide related predictive constructions (Shalizi & Crutchfield, 2001; Littman et al., 2001) (Appendix A). Finite measurements constrain only tested comparisons; tolerance-based matching need not be transitive and does not by itself identify these exact classes.

A deterministic scalar example separates the information requirements. In fixed physical units, assume the response class

$$z = g_p u, \qquad O = g_r z,$$

with zero baseline, no direct source–endpoint path, and calibrated healthy gains $g_p = g_r = 1$. The two changes $(g_p, g_r) = (1/2, 1)$ and $(1, 1/2)$ have the same complete source–endpoint map $O = u/2$. Before state-clamp or repair feedback, source interventions observed only at the endpoint cannot distinguish them. State measurements do distinguish $z = u/2$ from $z = u$. Alternatively, a known selective clamp $\mathrm{do}(z = v)$ gives $O = g_r v$ and identifies the readout gain without separately measuring the natural state.

An executable rule in this class is to estimate g_p from the source–state contrasts, estimate g_r from independent state clamps, and check the predicted source–endpoint response against the measured one. Compare both gains with the calibrated reference. With one component reset allowed, reset g_p when only g_p differs, reset g_r when only g_r differs, leave a healthy system alone, and abstain from a single-fault attribution when both differ. At $u = 1$ the correct reset changes O from $1/2$ to 1 in either example; the other reset leaves $O = 1/2$. This consequence is deduced from the equations after the action is chosen. Ordinary state-space identification with the same measurements recovers the same gains and decisions.

The assumptions matter: $z = u$, $O = 2z$ and $z = u$, $O = 2u$ give identical source/state/endpoint records although only the first endpoint depends on the state. A selective state clamp separates them.

A fit to source interventions alone therefore cannot certify mediation. Outside the declared class, a mismatch calls for competing mechanisms rather than a forced component label.

Decision value also depends on the readout. For $\delta z = \mathrm{diag}(1, 4)\delta u$ and $\delta O = (1/4, 1/100)\delta z$, positive unit source steps have state norms $1, 4$ but endpoint changes $1/4, 1/25$. For a fixed upper-boundary gap $1/4$, equal step costs and allowed steps of at most one, only the first axis reaches the boundary. Standard constrained control gives the same choice. The accompanying code checks these deductions using exact arithmetic, including zero-gain, healthy, multiple-change and direct-path controls. These are analytic checks, not measured AI fault or repair results.

B.2.5 Comparing realization-specific retained responses

Agreement of nested readouts need not imply agreement of other endpoint observables. A cross-realization comparison must therefore specify both the state correspondence and the endpoint map.

Retained response and endpoint readout. Let $\mathcal{Z}_i^{\mathrm{reach}} \subset Z_{f_i}$, $i = 1, 2$, be the retained states reached in two realizations, and let $J : \mathcal{Z}_1^{\mathrm{reach}} \to \mathcal{Z}_2^{\mathrm{reach}}$ be a specified comparison map. If $(P_{f_2}, Q_{f_2}) \circ J = (P_{f_1}, Q_{f_1})$, corresponding states $z_2 = Jz_1$ have the same reduced readouts (w, r) for every $z_1 \in \mathcal{Z}_1^{\mathrm{reach}}$ with $P_{f_1}(z_1) > 0$. Assume each endpoint factors through its retained coordinate, with maps $\bar{O}_{f_i,T}$ into a common declared readout space. Exact endpoint values agree on corresponding states if and only if $\bar{O}_{f_2,T} \circ J = \bar{O}_{f_1,T}$ there. If the reached sets are locally smooth and these maps are differentiable, first-order endpoint changes under $\delta z_2 = DJ|_{z_1}\delta z_1$ agree if and only if

$$D\bar{O}_{f_2,T}|_{Jz_1}\, DJ|_{z_1}\delta z_1 = D\bar{O}_{f_1,T}|_{z_1}\delta z_1$$

for every declared reached tangent direction δz_1.

Derivation. The reduced-readout identity follows by composition with J and division by the common positive outer readout. Exact endpoint agreement concerns the separate maps $\bar{O}_{f_i,T}$. The derivative condition is the chain rule applied to corresponding displacements. Exact composition implies this condition; equal derivatives alone do not imply equal endpoint values, since the maps may differ by an additive constant. For scalar endpoint values, agreement only of their declared ordering is weaker than equality of the values.

Interpretation. Preparation, retained dynamics and endpoint projection have distinct roles within each realization. Their composition under matched semantics is the reusable response construction, not equality of the realization-specific state coordinates. No comparison map J is assumed among the physical and AI realizations reported here. Corresponding changes can be compared only after a state comparison map or matched readout semantics has been declared. Such a comparison must distinguish changes in the retained dynamics from differences in endpoint projection: one readout may compress a change that another resolves.

B.3 Conditional well-posedness and coherence

The specified field–memory model has a unique mild trajectory on finite time horizons under the regularity assumptions below. Internally generated delayed feedback acts through the augmented history; externally prescribed delayed signals are supplied as input histories. The initial field, memory state, internal histories and prescribed inputs then determine the trajectory over the observation window. This is uniqueness of evolution from a complete preparation, not attraction of nearby trajectories.

For a finite maximum internal delay $\tau_{\max}$, take the dynamical base space $\mathcal{B}_{\mathrm{dyn}} = H \times \prod_i M_i$, with Banach memory spaces, and the phase space $\mathcal{X}_\tau = C([-\tau_{\max}, 0]; \mathcal{B}_{\mathrm{dyn}})$ with the supremum norm. Write $x(t) = (c_t, m_1(t), \ldots, m_N(t))$ and $\zeta_t(\theta) = x(t + \theta)$. The prescribed continuous preparation ζ_0 has endpoint $x(0)$. Internal quantities and their histories are specified functionals of this modeled history; any extra generator is included in $\mathcal{B}_{\mathrm{dyn}}$. Point evaluation is bounded in this phase space. Passive observation records are subsequently defined by their integrals when integrable, rather than assumed to belong to the dynamical phase space.

The Lyapunov–Krasovskii-type coherence estimate requires additional regularity and aggregate dissipativity conditions. Under those conditions, it bounds field deviations and memory disagreement

together with the specified input-history contribution. This bound is not a prerequisite for the general finite-window response or the readout-defined regions; stable-branch response and stability geometry use their separately stated hypotheses.

Well-posed delayed field system. For any finite horizon $[0, T]$:

1. L generates a strongly continuous contraction semigroup on H;
2. the combined source and memory functionals are defined on all of $\mathcal{X}_\tau$, are Lipschitz on each bounded history set and have at most linear growth in $\|\zeta\|_\infty$. Write their values as $J_i(\zeta)$ and $\Psi_i(\zeta, \eta_i)$, retaining all internal delay arguments. Global componentwise Lipschitz bounds suffice for finite compositions; local bounds require the separate growth condition;
3. the apparatus inputs η_i are strongly measurable and essentially bounded on finite horizons. Explicit time/input dependence of the functionals is strongly measurable at each fixed history, with Lipschitz and growth bounds uniform on finite horizons and bounded input sets;
4. each externally prescribed signal y_ℓ is supplied on $[-\tau_\ell, T]$, is strongly measurable there and essentially bounded on that interval; and
5. each $E_\ell : Y_\ell \to H$ is a bounded linear operator.

Conditional mild well-posedness. Under the "Well-posed delayed field system" assumptions above, (B.1) together with (B.2) admits a unique mild solution on $[0, T]$ for the compatible continuous preparation ζ_0, with external delayed signals treated as prescribed inputs. If the functionals are defined only on an open neighborhood, the conclusion holds only up to exit from that neighborhood unless continuation is separately established.

Derivation. Rewrite the closed field–memory system in variation-of-constants form and apply fixed-point continuation for semilinear delayed systems on a Banach space of continuous trajectories (Pazy, 1983; Henry, 1981; Hale & Verduyn Lunel, 1993; Fridman, 2014). In this construction the free propagator carries the initial field, and every source injected at time s is propagated to the later time t. Writing the field history on the interval $[-\tau_{\max}, T]$, the equation becomes

$$c_t = e^{tL} c_0 + \int_0^t e^{(t-s)L} \Big(\sum_i J_i(\zeta_s) + \sum_\ell E_\ell y_\ell(s - \tau_\ell) \Big)\, ds.$$

The memory components satisfy the integral equations

$$m_i(t) = m_i(0) + \int_0^t \Psi_i(\zeta_s, \eta_i(s))\, ds.$$

The measurability and growth assumptions make these Bochner integrals well-defined. Because L generates a contraction semigroup and the induced source and memory maps are locally Lipschitz with at most linear growth, the integral operator is locally contractive on a sufficiently small ball in the field–memory history space $C([-\tau_{\max}, T]; H \times M_1 \times \cdots \times M_N)$ for short time. Picard iteration therefore yields a unique local mild solution. If $r(t) = \sup_{-\tau_{\max} \le s \le t} \|x(s)\|$, the growth bounds give $r(t) \le r(0) + at + b \int_0^t r(s)\, ds$, hence $r(t) \le (r(0) + at) e^{bt}$ for constants $a, b \ge 0$ on the fixed horizon. With globally defined functionals and the bounded-set Lipschitz condition, this excludes finite-time blowup and permits continuation to $[0, T]$. It does not imply boundedness as $T \to \infty$ or classical differentiability.

Coherence functional. For this estimate, take the memory spaces M_i to be Hilbert spaces. Fix an exact reference solution, including a stationary operating point at fixed controls, and write c, m_i, y_ℓ below for deviations from it. A moving instantaneous equilibrium instead contributes its tracking residual as an additional forcing term and is outside the estimate stated here. In the deviation equation and its bound, J_i denotes the difference between the nonlinear operator source evaluated on the perturbed trajectory and on the reference trajectory, including each trajectory's retained-history dependence; it is not the original nonlinear source evaluated on the deviation alone. To measure whether those deviations remain bounded under delayed forcing, use the coarse coherence functional

$$C(t) = \|c_t\|_H^2 + \sum_{i=1}^{N} a_i \|m_i(t) - \Gamma_i c_t\|^2 + \sum_{\ell=1}^{N_\tau} b_\ell \int_{t-\tau_\ell}^{t} e^{-2\alpha(t-s)} \|y_\ell(s)\|^2 \, \mathrm{d}s. \tag{B.9}$$

Here $a_i, b_\ell > 0$, $\alpha > 0$ is the chosen exponential discount rate for the delay history, and $\Gamma_i : H \to M_i$ is a bounded linear comparison map from the field to the memory component used to measure field–memory disagreement. The three terms measure, respectively, the field deviation, the mismatch between memory and the mapped field, and the exponentially discounted energy stored in the recent delayed input history; the discount at rate 2α is the standard Lyapunov–Krasovskii weight for delay channels.

Bounded-coherence inequality. This is a conditional sufficient estimate, not a consequence of well-posedness. For sufficiently regular trajectories remaining in a bounded neighborhood, write $e_i = m_i - \Gamma_i c$ and assume that, for some $\lambda_0 > 0$ and $d_\ell \geq 0$,

$$\begin{aligned} &2\Big\langle c, Lc + \sum_i J_i + \sum_\ell E_\ell y_\ell(t-\tau_\ell)\Big\rangle \\ &\quad + 2\sum_i a_i \langle e_i, \dot{m}_i - \Gamma_i \dot{c}\rangle \\ &\leq -\lambda_0\Big(\|c\|_H^2 + \sum_i a_i \|e_i\|^2\Big) + \sum_\ell d_\ell \|y_\ell(t-\tau_\ell)\|^2. \end{aligned}$$

This aggregate inequality states that internal dissipation dominates the operator, memory and delayed couplings. If, in addition, $b_\ell e^{-2\alpha\tau_\ell} \geq d_\ell$ for every delay channel, then

$$\frac{\mathrm{d}}{\mathrm{d}t} C(t) \leq -\lambda C(t) + \sum_{\ell=1}^{N_\tau} b_\ell \|y_\ell(t)\|^2, \qquad \lambda = \min\{\lambda_0, 2\alpha\} > 0. \tag{B.10}$$

The estimate applies only while the stated regularity and aggregate dissipativity hypotheses hold.

Derivation. Leibniz differentiation gives, for each history term,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{t-\tau_\ell}^{t} e^{-2\alpha(t-s)}\|y_\ell(s)\|^2\,\mathrm{d}s = \|y_\ell(t)\|^2 - e^{-2\alpha\tau_\ell}\|y_\ell(t-\tau_\ell)\|^2 - 2\alpha\int_{t-\tau_\ell}^{t} e^{-2\alpha(t-s)}\|y_\ell(s)\|^2\,\mathrm{d}s.$$

Adding this identity to the aggregate dissipativity bound and using $b_\ell e^{-2\alpha\tau_\ell} \geq d_\ell$ absorbs every delayed boundary term and yields (B.10). This is the standard weighted Lyapunov–Krasovskii estimate (Hale & Verduyn Lunel, 1993; Fridman, 2014).

Interpretation. Equation (B.10) is an energy budget for deviations: internal dissipation absorbs the delayed boundary terms, while the current input supplies the remaining forcing. It bounds the coherence functional under the stated dissipativity assumptions and bounded forcing. It does not establish controllability, return to an exact endpoint, or the existence or cost of a repair policy.

The empirical recovery metrics are coarse summaries of the shared coordination state. Lag, normalized deficit area, contradiction-recovery fraction and terminal recovery fraction are finite-dimensional readouts after perturbation; they do not directly test inequality (B.10). The estimate provides one sufficient local boundedness condition for the theoretical field law, rather than an empirical conclusion extracted from those readouts.

C Readouts and nested operating criteria

A scalar or vector-valued readout specifies which state changes are observed. The general response construction places no restriction on its number of components. A nested pair is useful when one criterion measures continued operation and a stricter one measures endpoint correctness. The following specialization fixes their allowed values without assuming autonomous retained-state dynamics.

C.1 Nested observables and reduced variables

The following definitions distinguish continued workability from endpoint exactness. Their logical nesting fixes the allowed readout values; it places no bound on retained-state dimension.

Family-specific nested observables. Fix one admissible family f, one still-workable local operating point (z_*, u_*), and one declared outer/inner observable pair on the local coarse-state manifold

Z,

$$P_f, Q_f : Z \to [0,1], \qquad 0 \le Q_f(z) \le P_f(z) \le 1.$$

Here P_f represents the outer continuation condition and Q_f a stricter endpoint condition contained within it; pointwise numerical order alone does not establish this nesting. Both conditions are limited to the declared protocol and observation window.

Canonical reduced variables. Define

$$w_f := P_f, \qquad r_f := \frac{Q_f}{P_f} \quad \text{when } P_f > 0.$$

Thus w_f measures remaining workability, while r_f measures the inner readout relative to the outer one. It is a conditional exactness fraction only when both readouts are nested measures on the same population.

Comment. If one defines $s_f := Q_f$, then $s_f = w_f r_f$. The quantity s_f is therefore a derived convenience variable, not a canonical readout coordinate.

Readout conventions. The symbols P_f, Q_f denote normalized family-specific readouts in $[0,1]$ for logically nested endpoint criteria. They may be support probabilities, empirical fractions over the same population, scores normalized to preserve the ordering of the logically nested criteria, or binary indicators from a single realized trial. The ordering $0 \le Q_f \le P_f \le 1$ is structural for this pair. Distinct macroscopic reductions that happen to be numerically ordered on an interval, as in the flock, remain separate components of a more general endpoint map and do not inherit the nested-pair regime labels defined in Section D.1. A step indicator is not differentiated. Whenever a first-order endpoint response is used, the realization declares either differentiable local extensions $\widetilde{P}_f, \widetilde{Q}_f$, a smooth endpoint vector O_f^{soft}, or signed score margins whose zero sets calibrate the binary boundaries. The boundary labels remain those defined by P_f, Q_f; the differentiable map supplies the response covectors.

A state coordinate invisible to both P_f and Q_f under admissible perturbations is irrelevant to this boundary classification. A quantity with an independent endpoint meaning instead requires a separate diagnostic or an enlarged readout declared before measurement; two components are not required by the general response theory.

For example, a differentiable readout that supplements the nested pair is

$$O_f^{\mathrm{soft}} = (\widetilde{P}_f, \widetilde{Q}_f, S_{1,f}, \ldots, S_{m,f}),$$

and its local response is obtained from $\mathrm{D}_z O_f^{\mathrm{soft}}$. A general readout need not contain this pair. For the displayed example, the boundary regime construction in Section D.1 is the projection induced by the calibrated nested pair. Additional observables define additional measured directions specified alongside, rather than after, the two-dimensional projection.

Representation sufficiency. For deterministic readouts, exactly equal retained records with different readouts under the same permitted continuation reject the representation. Stochastic readouts require a resolved difference in the declared response statistic or conditional readout law, with specified uncertainty control (Section B.2.3). Repairing the representation may require re-encoding within the same dimension or changing its coordinates. The failure concerns the representation, not logical nesting. Conversely, a proposed extra coordinate adds no distinction for this family if it is invisible to $\mathrm{D}_z O_f^{\mathrm{soft}}$, leaves calibrated boundary margins unchanged, and preserves the binary verifier under finite interventions throughout the admissible neighborhood.

D Readout boundaries and susceptibility

Nested observables define regions and boundaries in the readout space. Loading-path assumptions determine how those boundaries are encountered. Finite-window response describes how sources change the readouts; stationary susceptibilities require the additional closure and stability conditions of Section B.2.1. The same measurement organization does not impose a shared response curve across realizations.

D.1 Local readout regimes, boundary surfaces, and susceptibility

The nested readouts define regimes. Separate path assumptions determine whether and in what order their boundaries are encountered.

Admissible loading path. Fix one admissible family f and let U_{adm} be a local control manifold on which the regular operating branch $u \mapsto z_*(u)$ exists smoothly. Call a C^1 path $\gamma : [0,T] \to U_{\mathrm{adm}}$ an *admissible loading path* when it keeps the same instance class and the same outer/inner observables fixed, and when its readout sequence has no re-entry: after leaving $r = 1$ it never returns to $r = 1$, after reaching $r = 0$ it never returns to $r > 0$, and after reaching $w = 0$ it remains failed. These conditions define monotone loading but do not impose the relative locations of the first-hit thresholds. The path must be fixed by intervention semantics before the crossing data are read; a path chosen after observing the transition order is not an admissible loading path.

Local readout regime structure. For any admissible family with canonical reduced variables

$$w(u) := P_f(z_*(u)), \qquad r(u) := \frac{Q_f(z_*(u))}{P_f(z_*(u))} \quad \text{when } P_f(z_*(u)) > 0,$$

the allowed local readout regimes are

$$\text{exact stratum} : w > 0,\ r = 1, \qquad \text{partially exact region} : w > 0,\ 0 < r < 1,$$
$$\text{workable-but-inexact stratum} : w > 0,\ r = 0, \qquad \text{failed stratum} : w = 0.$$

Local boundary surfaces. Define the four readout strata

$$S_{\mathrm{E}} := \{u \in U_{\mathrm{adm}} :\ w(u) > 0,\ r(u) = 1\}, \qquad S_{\mathrm{PE}} := \{u \in U_{\mathrm{adm}} :\ w(u) > 0,\ 0 < r(u) < 1\},$$
$$S_{\mathrm{W}} := \{u \in U_{\mathrm{adm}} :\ w(u) > 0,\ r(u) = 0\}, \qquad S_{\mathrm{F}} := \{u \in U_{\mathrm{adm}} :\ w(u) = 0\},$$

the exact, partially exact, workable-but-inexact, and failed strata. Define each transition locus as the common relative boundary of the two adjacent strata,

$$\Sigma_{EW} := \overline{S_{\mathrm{E}}} \cap \overline{S_{\mathrm{PE}}}, \qquad \Sigma_{WR} := \overline{S_{\mathrm{PE}}} \cap \overline{S_{\mathrm{W}}}, \qquad \Sigma_{RO} := \overline{S_{\mathrm{E}} \cup S_{\mathrm{PE}} \cup S_{\mathrm{W}}} \cap \overline{S_{\mathrm{F}}}.$$

The closures for Σ_{EW} and Σ_{WR} are taken relative to $U_+ = \{w > 0\}$, and the closure for Σ_{RO} relative to U_{adm}. Assume each locus is locally a C^1 codimension-one hypersurface in this relative topology and the loading path meets it transversally at nonzero rate. At each encounter, additionally require the path to pass between the two defining strata on nonempty adjoining intervals. This is an actual change of regime, not merely an intersection with their closures; for Σ_{RO}, the two sides are workable and failed. These assumptions can fail at plateau edges, intersections or corners. The set identities alone establish neither smoothness, actual regime changes nor encounter order. The parameters s_{EW}, s_{WR} and s_{RO} are the first encounters with the corresponding transition loci along the declared path; an unencountered locus has first-hit parameter $+\infty$. The four-step statement requires all three encounters within the observation interval. Along a continuous admissible monotone path starting in the exact stratum, the following order holds only when the first-hit parameters satisfy $s_{EW} < s_{WR} < s_{RO}$:

exact stratum $\to$ partially exact region $\to$ workable-but-inexact stratum $\to$ failed stratum.

Derivation. Step 1 (Hypersurface structure). Nesting gives $0 \le r \le 1$ on U_+. Failure is specified by $w = 0$, where r is undefined; for discontinuous readouts, membership of a threshold point follows the declared equality convention. Smooth common relative boundaries and transverse encounters that change the adjacent strata are assumed, not derived from nesting.

Step 2 (Monotonicity excludes re-entry). The path definition excludes return to any departed regime. Actual crossings supply nonempty intervals in the adjacent strata. With all three finite encounters ordered as $s_{EW} < s_{WR} < s_{RO}$, these assumptions yield the displayed sequence. Without separation, the path can pass directly from partial exactness to failure.

Step 3 (Exceptions). Coincident thresholds can also skip an intermediate regime; they are nongeneric only if the threshold conditions vary independently. Regularity and transversality give isolated encounters, but do not enforce their order. The separated sequence remains a family- and path-specific hypothesis.

Boundary-test selection principle. For a fixed instance class, select controls whose projected responses distinguish exactness loss from workability loss while preserving admissible monotonicity along the same continuation.

Comment. The nested readouts fix the admissible regimes. Boundary regularity and the actual changes of regime and first-hit order are separate local assumptions. Threshold locations, curve shapes, and the placement of particular family instances relative to those boundaries are empirical calibration questions.

D.2 Susceptibility and comparison of macroscopic readouts

The primary object is the finite-window response from a declared source perturbation, through the retained state, to an endpoint readout. Continuum kernels and stable-branch susceptibilities are related limits, not alternative names for the same measurement.

Susceptibility kernel. For a perturbation $h_\beta(\mathbf{y}, t')$ applied to channel β, the continuum field susceptibility is

$$\chi_{\alpha\beta}(\mathbf{x}, \mathbf{y}; t, t') = \frac{\delta\langle c_\alpha(\mathbf{x}, t)\rangle}{\delta h_\beta(\mathbf{y}, t')}.$$

For a deterministic finite graph G, or a pathwise derivative with the noise realization held fixed, the response operator and its readout contraction are

$$\widehat{\chi}^G_{\alpha\beta}(x, y; t, t') = \frac{\partial c_\alpha(x, t)}{\partial h_\beta(y, t')}, \qquad \widehat{\chi}^{\mathrm{red}}_{aj} = \frac{\partial O_a}{\partial \theta_j}.$$

Here $x, y \in V_G$ are vertices of $G = (V_G, E_G)$, t, t' are discrete times, and α, β index the declared finite channel basis; $c_\alpha(x, t)$ is the field-channel value. The amplitude $h_\beta(y, t')$ multiplies a unit source direction at the specified vertex, channel and time. O_a is a smooth readout component, and θ_j specifies a source-control direction. Applying these source directions and readout differentials contracts the finite field operator to $\widehat{\chi}^{\mathrm{red}}$. This is a discretized susceptibility estimate; identification with a continuum kernel requires graph, time and channel refinement, and identification with χ_* requires the stable limit specified below. For stochastic mean responses, the finite-graph counterpart of the continuum kernel differentiates the ensemble mean. It equals the mean of pathwise derivatives only when differentiation and expectation may be interchanged. A nonlinear ensemble readout requires the additional state-response information specified below. The derivative construction requires the finite-window solution map, retained-state projection and readout to be differentiable on that class. Local Lipschitz continuity ensures well-posed evolution under the preceding assumptions but does not imply this differentiability. At nonsmooth operating points, finite contrasts or specified one-sided responses replace the two-sided linear derivative. At fixed horizon T, suppose the endpoint response factors differentiably through $z_T = G^{\mathrm{cg}}_f(s_T)$ on that source class. For a discrete realization, A_t propagates an existing augmented-state displacement through one update, and B_t maps a source perturbation into that state. Linearization about the fixed reference protocol gives $\delta s_t = A_t \delta s_{t-1} + B_t \delta u_t$; zero initial displacement then gives

$$\delta O_T = \mathrm{D}_z \bar{O}_T \, \mathrm{D}_s G^{\mathrm{cg}}_f(s_T) \sum_{t \le T} \Phi(T, t) B_t \delta u_t, \qquad \Phi(T, t) = A_T \cdots A_{t+1}.$$

If the implemented graph already contains every required memory coordinate and is itself the declared retained state, the coarse-graining derivative is the identity. Otherwise it must remain in the source-to-state-to-readout chain.

For the time-summed retained response, suppose the graph itself is the retained state or admits the verified closed transition above. Write A, B for its time-independent retained transition and source maps. When $\rho(A) < 1$, a persistent tangent source gives

$$\chi^G_* = (I - A)^{-1} B = \sum_{k=0}^{\infty} A^k B.$$

A one-time source at t' instead gives the retarded Green function $A^{t-t'} B$ for $t \ge t'$ and zero response before t'. This discrete resolvent acts on retained coordinates. For a general full-state resolvent, the

projection differential must first be applied to obtain retained-state response. The retained resolvent is analogous to the continuum branch response

$$\chi_* = -[\mathrm{D}_z f]^{-1}\mathrm{D}_u f,$$

with equality requiring a justified identification, such as the exact zero-order-hold correspondence in Eqs. (B.6)–(B.8), or a demonstrated graph-and-time refinement limit.

The constructed code-repair graph checks these finite-dimensional identities: its assembled branch has $\rho(A) = 0.941$, while the reported branch residual is only a back-substitution check of the solve for assembled A and B. The binary verifier is evaluated through a smooth score surrogate for derivatives and through fixed score margins for the declared binary boundaries. Because endpoint information enters the graph construction, this is an implementation check rather than a measurement of language-model dynamics. Section E.2 reports the fixed-record replay and fresh model-readout experiments, including their controls and scope.

The coupling family used as a spectrum control is likewise constructed rather than identified from data. With graph diffusion and channel coupling, its stable-branch response can range from a broad spectrum to concentration on one dominant mode as coupling approaches the stability boundary. For the declared baseline retention factor $\rho_0 = 0.6$ of the constructed propagator, distinct from its assembled spectral radius $\rho(A)$, the participation ratio changes from 32 in the decoupled limit through about 24 at intermediate coupling toward 1 near the stability boundary. It illustrates how coupling redistributes response; it is not fitted to the code-repair records and supplies no cross-realization rank law.

A continuum interpretation requires more than a fine graph. Let $h_n = \Delta t_n \to 0$, let $\rho(A_n) < 1$, and let B_n be a per-unit-time graph source operator, rather than an already time-integrated source step. Let I_n map graph states to continuum states and P_n^U map the declared source space to graph inputs. Let $\mathcal{L}$ generate the complete linearized continuum evolution approximated by A_n, on the corresponding state and history space. It includes the source, memory and feedback derivatives and is distinct from the autonomous medium generator L in Eq. (B.1) unless these contributions vanish. The source operator and interpolation maps act on the same declared spaces. Assume that $\mathcal{L}$ generates a strongly continuous semigroup, 0 is in its resolvent set, and B is bounded. Sufficient conditions, in operator norm on that source space, are joint source-propagator consistency for each $t \geq 0$ and a uniform bound for all n, k:

$$I_n A_n^{\lfloor t/h_n \rfloor} B_n P_n^U \to e^{t\mathcal{L}} B, \qquad \|I_n A_n^k B_n P_n^U\| \leq C e^{-\omega k h_n}, \qquad C, \omega > 0.$$

The identity

$$h_n \sum_{k\geq 0} I_n A_n^k B_n P_n^U = \int_0^\infty I_n A_n^{\lfloor t/h_n \rfloor} B_n P_n^U \, \mathrm{d}t$$

and dominated convergence then give

$$I_n (I - A_n)^{-1} h_n B_n P_n^U \to (-\mathcal{L})^{-1} B.$$

The joint conditions control source components that can propagate through coordinates hidden by projection; separate projected consistency of A_n and B_n does not suffice. These are refinement conditions, not claims that a reported finite graph realizes the continuum limit.

Matched susceptibility and readout response. For matched perturbation channels and the same fixed linear readout, equal mean-field susceptibilities give equal first-order mean endpoint responses. The same statement holds for smooth readouts linearized about a deterministic reference with the same readout differential. Equality of the susceptibility alone does not imply agreement under different readout maps.

Derivation. For a fixed linear coarse observable $O[c](t) = \int K_\alpha(\mathbf{x}, t) c_\alpha(\mathbf{x}, t)\, d\mathbf{x}$,

$$\delta\langle O(t)\rangle = \int K_\alpha(\mathbf{x}, t)\chi_{\alpha\beta}(\mathbf{x}, \mathbf{y}; t, t')\delta h_\beta(\mathbf{y}, t')\, d\mathbf{x}\, d\mathbf{y}\, dt'.$$

The conclusion follows by contraction with the common readout kernel. A smooth readout linearized about a deterministic reference has the same first-order form with $K = \mathrm{D}O|_{c_*}$. For nonlinear observables of a fluctuating field, the ensemble response requires a separately justified closure or the joint state and response statistics. The statement is conditional on matched channels, coarse-graining and readout semantics; none of the empirical realizations is claimed to share a measured susceptibility, response spectrum, dimension or boundary geometry.

E EMPIRICAL REALIZATIONS AND MEASUREMENTS

Table 7 distinguishes the empirical and constructed tests of the response protocol. Each detailed protocol reports its positive and negative outcomes. Only nested readout families enter the (w, r) classification; no common state coordinate or empirical response law is assumed.

Table 7: **Evidence roles in the response framework.** A rejected state description can be a successful diagnostic result. Constructed controls, numerical calibrations and fresh model measurements support different inferences.

Evidence	Result	Scope
Analytic response and mechanism controls	Kernel inclusion characterizes first-order factorization; state measurements separate propagation and readout gains.	Conditional mathematical deductions. Response agreement alone does not identify mediation (Appendix B.2.4).
Delayed consensus	Measured amplitudes and phases agree with analytic predictions.	Calibration of a specified driven system; no cross-system empirical response law (Section E.4.1).
Polar flock	Nonlinear steady-state readouts first fall below one half at different sampled noise levels.	Nonnested observables, with threshold-dependent numerical order (Section E.4.2).
Constructed graph and record replay	Composed and directly computed responses agree; recorded source events populate the interface.	Implementation checks; replay reprocesses fixed records under prescribed graph dynamics (Section E.2).
Fresh model readouts	Two of twelve directions pass step refinement; the all-direction criterion fails.	Measured readout response after imposed graph propagation. Linearly independent terminal-state contrasts and restricted model inputs leave retained-state sufficiency untested (Section E.2).
State-description counterexamples	An archived worksheet comparison and a constructed raw-sensor bypass reject different proposed states.	Counterexamples within their permitted-input classes, not evidence of a generally sufficient graph (Section E.2.1).
Prospective diagnostic pilot	One component is corrected; complete repair is $0/2$ faulty programs and joint correctness remains $2/4$.	New settings under fixed rules; the positive criterion fails (Section E.2.2).
Delayed-tool intervention	A preselected answer-free join corrects one repeated failure. A separate 24-call follow-up restores $14 \to 9$ in all three fresh pairs; raw input, checklist and unrelated-note controls remain $9 \to 9$.	Exploratory, same-case evidence after failed calibration, not held-out validation. Ordinary diagnosis agrees and fixed joining is cheaper (Section E.5).
Evidence-selection development study	The primary policy and direct reading have identical correctness across 178 within-task panels; the primary policy costs more requests.	Descriptive analysis of 89 complete tasks from a 96-task roster, with incomplete state-measurement coverage and reused tuning folds; no independent confirmation (Section E.6).
Persistence conditions	None of four alternatives matches the persistent reference on every recovery and load coordinate.	One task, point estimates only; incomplete source and trajectory-rule provenance prevent a necessity claim (Section E.2.4).

E.1 AI COORDINATION FAMILIES AND EMPIRICAL BASIS

Each nested-readout family fixes an instance class, interface, stopping rule, loading control and logically nested workability/exactness criteria, as in Section 3. Four task families supply the recorded comparisons:

- *Constraint allocation*: local workability and exactness under a fixed constraint class; constraint splitting or round thinning varies coordination load.
- *File-grounded tool use*: grounded tool-path workability and exact answers; tool-budget or file-grounded-budget continuation varies access to delayed evidence.
- *Code repair*: verifier-valid repair and full-suite correctness in the upstream workflow. The endpoint-informed graph check instead tests action selection and target agreement without reexecuting the test suite.
- *Rule-gated counting*: contradiction-sensitive continuation and exact counting; contradiction density or delayed-evidence stress varies retention load.

The instance class and readouts remain fixed along each loading path. The graph implementation, source-construction replay and fresh model readout experiments have different evidence roles, specified below.

E.2 EMPIRICAL PROTOCOLS AND CONSTRAINED COMPARISONS

Response operators and numerical checks. Table 6 specifies the local source mechanisms; Section D.2 defines the channel-resolved finite-graph response, source amplitudes and stochastic derivative conventions. For the finite-window calculations, A and B follow from the declared dynamics or finite source perturbations, rather than a fit to endpoint scores. Delayed consensus uses delay-augmented local equations. In the code-repair example, the graph is constructed using endpoint information, and the endpoint readout uses its scalar values and type labels. Its source basis separates source classes by site and channel. These choices establish neither minimality, state sufficiency nor an internal language-model state dimension.

Finite-window responses use the transition products in Eq. (2). For a closed retained-graph linearization with time-independent maps A, B and $\rho(A) < 1$, the time-summed response is obtained from $(I - A)\chi_*^G = B$. The residual $\|(I - A)\chi_*^G - B\|_\infty$ checks only that solve for the assembled matrices. The constructed coupling control uses $r_{\mathrm{PR}} = (\sum_k \sigma_k^2)^2 / \sum_k \sigma_k^4$ for the singular values σ_k of its nonzero response matrix. Consensus instead bounds $r_{\mathrm{eff}} = (\sum_k E_k)^2 / \sum_k E_k^2$ with $E_0 = \mathcal{A}^2$ and $\sum_{k\geq 1} E_k = V$. The bounds and conversion from sample-covariance to quadrature weights are given in "Reduction accuracy" (Section E.4.1).

The finite-window protocols keep the measurement conditions fixed while perturbing either the shared medium or a source coordinate. The flock instead varies its steady-state noise loading under fixed readout definitions. Within each comparison, the instance class, interface, stopping rule where applicable, loading semantics and observable pair remain fixed.

For the regime comparison, the instance and model classes, interface, tool depth and readout are fixed while one loading coordinate varies. The stated boundary-ordering criterion specifies the expected next crossing under its hypotheses. Only paths retained by the inclusion rule enter this retrospective comparison; it is not prospective. Persistence interventions compare intact carryover, removed carryover, hidden contradiction carryover and restored carryover, holding fixed the underlying model family, task semantics, answer format, admissible action set, interface, stopping rule and endpoint readout. Under the same declared execution envelope and readout language, trace visibility, visible candidate count, contradiction handling and the coherence summary place runs in the nested readout coordinates (P_f, Q_f).

Selected-path records and missing provenance. The seven retrospectively selected paths cover three author-constructed file-grounded spreadsheet tasks in the typed-tool interface. They supply pooled endpoint coordinates, not aggregate benchmark scores. Source records enter through the attached grounding file declared for each instance. Available fields identify the task and its level, grounding-file name and type, parsed answer and response record, correctness and tool use, tool-step metadata and run index. Correctness and tool use determine the declared Q_f and P_f, placing a

run in (P_f, Q_f) or (w, r). The tool-step budget is swept across the included records. Elapsed-time and message-count fields record elapsed time and language-model-call count.

No dedicated fields record the declared stopping rule, an independent pre-scoring timestamp, temperature, seed or model revision. Per-run tool-call and verification-update counts are also absent. Tool-step metadata and run indices cannot recover these missing counts or other provenance. The records therefore do not establish that the protocol was fixed before scoring or permit complete reproduction. A faithful rerun must preserve the admissible range, instance definition, readouts, interface, execution envelope and loading semantics. The records document preparation and endpoint coordinates without reconstructing a source-to-retained-state response (Section E.2.3).

The same finite-window response calculation is implemented on a typed-evidence finite graph. The graph is constructed using endpoint information, and its linearized local branch is represented on a finite state space. At the selected operating point, the graph has spectral radius $\rho(A) = 0.941$, the largest magnitude of an eigenvalue of its one-step propagator A. Because $\rho(A) < 1$, the linearized dynamics are stable and the resolvent $(I - A)^{-1}$ exists; the proximity to unity gives a slowly decaying mode. The branch residual is $\|(I - A)\chi_*^G - B\|_\infty = 1.55 \times 10^{-15}$. In a separate finite-window propagation check, direct finite differences and the assembled transition products agree to maximum absolute error 1.65×10^{-13}. Both are implementation checks of the constructed dynamics, rather than empirical AI response laws. The binary verifier readout Q_f is an exactness indicator and has no ordinary differential at its jump. Endpoint response is therefore evaluated with the smooth surrogate

$$\boldsymbol{q}_{\mathrm{soft}}(z) = \mathrm{softmax}\big(s(z)\big), \qquad q_{\mathrm{soft}}(z) = \big[\boldsymbol{q}_{\mathrm{soft}}(z)\big]_{a_*},$$

where $s(z)$ is the four-action score vector, distinct from the augmented state $s(t)$, and a_* is the expected repair action. The readout recommends the highest-scoring action only when its score reaches the recorded threshold; otherwise it emits no action. Writing a_{rec} for that recommendation, the recorded verifier sets $Q_f = \mathbf{1}[a_{\mathrm{rec}} = a_*]$. Since a_* belongs to the declared action set and P_f records emission of an action from that set, this implementation satisfies $Q_f \le P_f$. The smooth probability is a response surrogate, not the binary boundary coordinate. The calibration to the binary readouts is made in raw-score space:

$$m_P(z) = \max_a s_a(z) - \theta, \qquad m_Q(z) = s_{a_*}(z) - \max_{a \neq a_*} s_a(z),$$

where θ is the recorded readout threshold. Away from action ties, $P_f = 1$ exactly when $m_P \ge 0$, and $Q_f = 1$ exactly when $m_P \ge 0$ and $m_Q > 0$; ties follow the recorded action-order convention. Thus $\mathrm{D}_z \boldsymbol{q}_{\mathrm{soft}}$ supplies a differentiable endpoint response, while m_P and m_Q locate the declared binary boundaries. The labeled expected repair is used to define parts of this constructed graph, including action labels, hint and test routing, and some source weights. Consequently, this is an implementation check of the finite-graph calculation on a graph constructed using endpoint information, not a measurement of language-model dynamics or a no-leakage AI experiment.

The deterministic implementation-level check perturbs the four constructed source-event amplitudes in each dry-run task. Their event site, source class, and event time define a source embedding E_{src} into the 8-site, 4-channel graph state. With four endpoint probabilities, this gives 16 derivative components per task. Direct central finite differences of the raw action scores agree with the corresponding derivatives obtained by composing the state response with the raw-score readout differential to relative error 5.33×10^{-14}. After the softmax projection, direct endpoint finite differences agree with $\mathrm{D}_z \boldsymbol{q}_{\mathrm{soft}}\, A^{t-t'} B E_{\mathrm{src}}$ to maximum absolute error 1.13×10^{-8}, relative ℓ_2 error 1.10×10^{-7}, and coefficient of determination $R^2 > 1 - 10^{-13}$. The central-difference step is 0.003 in source-amplitude units. Increasing it to 0.3 changes the finite-difference estimate by at most 1.10×10^{-3} relative to the baseline finite-difference estimate. A no-propagation structural ablation has relative error 1.00 and $R^2 = 0$; a cyclically permuted endpoint readout has relative error 1.63 and negative R^2. These deliberately altered calculations check that the implementation uses the declared propagation and endpoint map. Their numerical values largely reflect, respectively, the zero-propagation definition and the cyclic readout permutation, so they are implementation controls rather than independent empirical validation. The two code-repair tasks repeat the calculation with the same graph template and separately labeled finite operators, so they do not constitute independent graph-topology replications.

A separate source-provenance check begins with typed records generated by a live Qwen2.5-14B-Instruct-MLX run on seven file-grounded tasks. The recorded run produced usable answer candidates for all seven tasks and exact candidates for six. Each task contributes the three source records

actually present in the recorded run—plan, tool result and answer candidate—with no synthetic verification event added when that record is absent. The graph action set is derived from the visible answer candidate. Source amplitudes are assigned from visible planning and tool-status records together with a fixed answer-candidate event rule; the hidden expected answer enters only after readout in the endpoint verifier. For the single nonexact task, that hidden correct answer is absent from the action set and from every source event. Replaying the resulting 21 typed source events through the deterministic finite-graph response layer uses the same source amplitudes $(0.7, 1.0, 1.0)$ for all seven tasks because their recorded planning and tool statuses coincide. They are seven provenance records passed through one numerical response template, not seven independent response realizations. The three declared source classes bound the source-channel rank by three. For the declared channel coupling and state representation, the stored nonzero response matrix has algebraic rank three; the value is therefore a property of the complete replay construction, not a consequence of the number of source classes alone. Probability normalization bounds the two-action decision-probability tangent rank by one; the stored nonzero tangent response attains rank one. The two-time kernel is zero before each source time, and its maximum spread across finite-difference amplitudes, 9.71×10^{-14}, is a numerical self-consistency check. The live run establishes the provenance of the source records; the response calculation remains a deterministic replay of fixed records. It neither identifies internal language-model dynamics nor perturbs and reruns the live model.

Fresh model readouts after imposed propagation. A separate protocol applies paired source-amplitude interventions before propagation and queries Qwen2.5-14B-Instruct in its MLX 4-bit quantized implementation afresh for every arm. The visible data contain only task text, allowed values and evidence records; endpoint verifiers are stored in a separate file. The runner first executes all source-extraction blocks. For each block it normalizes the source events, constructs the finite retained graph, binds the visible-task, extractor-request, extractor-response, event and graph hashes into a source seal, and appends that seal with an explicit filesystem synchronization. Only after all four confirmatory source seals are durable does the process open the verifier file and begin endpoint calls. The publication record is bound to a clean Git tree, the served model identifier and path, individual model-file hashes, code and data hashes, all API request and response identifiers, and the complete planned block denominator.

The seven semantic instances share one synthetic three-field constraint-planning template. The first three instances and two requested generation seeds form a development pilot; the remaining four instances and one requested seed form the confirmatory measurement and are excluded from protocol calibration. The endpoint operating load $\lambda = 0.9375$ was selected during development to leave the saturated exact-output regime, and is not an independently calibrated natural boundary. The endpoint input for each field is λ times its mean retained support plus $1 - \lambda$ times the candidate-specific support; at the selected load, within-field contrast is therefore attenuated by a factor of 16 before scoring. The same load is used for every confirmatory coordinate. The pilot fixes a minimum Euclidean terminal-state distance 0.05 and minimum absolute candidate-support slope 0.01 as analysis thresholds before the confirmatory calls. The retained graph and its deterministic propagation rule are externally constructed measurement coordinates rather than internal language-model state. Propagation uses retention 0.72, diffusion 0.45, and four observation steps, with each event deposited at its assigned time before that step's update. An update adds 0.72 times the deposited state to 0.45 times its graph-transported value, rescales by its maximum if that maximum exceeds one, then clips coordinates to $[0, 1]$.

Every source event has baseline amplitude 0.5. For each selected source direction d, paired interventions change only that event's amplitude by $\pm\epsilon$, with $\epsilon = 0.1$; refinement uses $\epsilon/2 = 0.05$. The imposed terminal-state contrast and measured endpoint finite-difference slope are

$$\Delta z_d(T) = z_{d,+}(T) - z_{d,-}(T), \qquad \widehat{g}^s_d(\epsilon) = \frac{s_{d,+} - s_{d,-}}{2\epsilon}.$$

The terminal-state distance is $\|\Delta z_d(T)\|_2$, over all coordinates of the declared terminal graph field. Here s is a scalar readout score, distinct from the augmented state $s(t)$. Each candidate is queried separately with a binary Y/N prompt. Within each field, its conditional Y likelihood is divided by the sum of those likelihoods over all allowed values. The score s is the arithmetic mean, over the three fields, of the normalized support assigned to each verifier-selected value. The verifier selects these values only for endpoint evaluation, after the source seals. Each prompt contains the complete numerical retained-support table and one candidate label, but no raw evidence record, and asks

whether that candidate is best supported. The endpoint scorer therefore implements a model-based readout of the declared graph rather than an independent planning step. Source extraction uses temperature 0; endpoint scoring uses temperature 1.0 and top-$p = 1.0$. Requested seeds derived deterministically from the base generation seed 4101 and candidate index are matched across paired counterparts. This construction is not a calibrated joint choice probability. Exactness Q is recorded after the calls and is not differentiated; workability is not varied in this protocol.

Every selected block remains in the denominator, with no adaptive retry, parser repair or result-dependent exclusion. All 4 confirmatory blocks complete, comprising 4 source-extraction calls and 648 endpoint-scoring calls (652 fresh live calls in total). Of the endpoint calls, 432 score the four perturbation arms across the 12 selected source directions; 216 are repeated-input and routing-control calls. The extractor produces 15 source events across the four blocks. Before endpoint unblinding, a fixed endpoint-blind ordering by event time, type, document, field, value and identifier selects the first three events in each block, leaving 3 events unprobed. By construction, every selected direction exceeds the terminal-state threshold, and the 12 directions yield only five distinct displacement magnitudes under the deterministic graph rule. Nine directions exceed the endpoint threshold, whereas all three directions in the reactor-inspection instance remain below it. These are retained-graph displacements whose endpoint slopes remain below the declared analysis threshold, not zero readout responses. A descriptive post-run instrument diagnostic finds that in 547 of the 648 endpoint calls (84.4%), the larger conditional Y/N probability is at least 0.99; the corresponding count is $355/432$ (82.2%) on the perturbation arms. Saturation may contribute to endpoint insensitivity, but these descriptive counts do not separate its effect from other properties of the readout. Only 2 of 12 directions satisfy the fixed slope-agreement test

$$\frac{|\widehat{g}_d^s(\epsilon) - \widehat{g}_d^s(\epsilon/2)|}{\max\{|\widehat{g}_d^s(\epsilon)|, |\widehat{g}_d^s(\epsilon/2)|, 10^{-12}\}} \leq 0.25.$$

The numerical floor fixes the normalization near zero. Four of the nine above-threshold directions reverse sign under this step refinement, and only five are both above threshold and sign-stable. The all-direction endpoint criterion and the local-linearity criterion are not met, so the experiment does not establish a differential susceptibility across the tested directions. These counts are descriptive for the twelve selected directions and carry no population-level null interpretation. The tested steps are 20% and 10% of the baseline source amplitude, so the failed linearity check does not distinguish a step outside the local regime from nonlinear or discrete readout behavior.

Within each instance, the three imposed terminal-state contrast vectors write to distinct per-event graph coordinates and are therefore linearly independent in the declared 21- or 22-component state. Their three-column finite-contrast matrix has zero kernel, so these columns supply no state-null combination for a first-order test. This finite-contrast statement does not identify the differential map $\mathcal{K}_{z,T}$ or establish retained-state sufficiency. The three unprobed events comprise two late coupling events and one support event; their omission also prevents a claim about the complete recorded source space.

The two identical-input endpoint calls returned identical outputs. State restoration and removal of graph propagation return zero endpoint difference because the software routing is clamped by construction; they are interface checks rather than empirical evidence of mediation. The four confirmatory instances are semantic variations of one template, so no task-family confidence interval or generality claim is made. The experiment does not measure a workability-boundary crossing, live multi-agent dynamics or a natural unloaded operating point. The endpoint prompt receives only the task text, the complete numerical retained-state representation and one candidate label, not the raw evidence records; retained-state sufficiency is therefore imposed by this protocol rather than tested by it.

E.2.1 Offline graph diagnosis and tool repair

A source response can suggest where to intervene without guaranteeing that the intervention repairs the endpoint. A separate offline example connects these two questions using the same typed-event graph engine as the AI response calculations. It wraps two archived model-written arithmetic programs in a new constructed apparatus, with nine constructed controls. The original program bytes are fixed; no new model is called. The graph is part of the new execution path, not a description of how the original model generated its program. The two programs are model-written baseline

outputs from an earlier 30-call development experiment comprising 6 preparations and 24 continuations. None of its 18 eligible accuracy arms passed. Arithmetic and formula defects had been identified before these two cases were selected and the diagnostic rule was defined. The cases and rule are therefore retrospective. The new offline tool interventions do not reverse the earlier failure, and the original raw calculations remain unchanged.

Inputs and retained state. Each task has paired sensor readings at three positions, preparation values computed from their means, and a separately saved worksheet. Preparation and worksheet values are distinct inputs, even when their numerical values agree. The flow task uses rates $(2, 4, 3)$ over three unit-duration intervals and asks for total volume and the time at which cumulative volume reaches 4.5. The calibration task uses values $(10, 20, 40)$ at coordinates $(0, 1, 2)$, with linear interpolation, and asks for the integrated response and the coordinate at which the response reaches 15. Units and endpoint meanings are task-specific.

Six events, three from each input stream, enter twelve graph sites with two typed channels. Values in the declared range $[0, 64]$ are divided by 64 and injected at tick zero. The positive event polarity denotes a numerical write, not endpoint correctness or confidence. Forward edge weight 1, return weight 0.15, retention $1/2$, propagation coefficient $1/4$ and zero channel coupling are fixed. Both ticks, 0 and 1, are stored. The candidate retained state is the full terminal graph field. Its readout selects six action/channel coordinates and multiplies them by 256: an input v gives terminal action value $v/256$, so the decoded input map is an analytically calibrated identity. This construction tests execution of the graph interface, not emergent coordination or a learned state. The saved field is reloaded before either the original program or a replacement tool reads preparation and worksheet values. Raw sensor history, target and interval durations (flow) or coordinate spacings (calibration) remain separate permitted inputs; they are not silently identified with graph coordinates.

Response checks and decision rule. Paired source probes add ± 0.1 to both sensors at every position, then recompute preparation and worksheet with any declared preparation offset held fixed. The diagnostic checks three task relations: preparation equals the sensor mean; the inverse answer lies within the task interval; and the total's central-difference slope under this uniform offset equals the interval length (3 for flow, 2 for calibration). The comparison tolerance is 10^{-8}. Exactly one violated relation selects, respectively, recomputation of preparation, replacement of the inverse calculation, or replacement of the integral calculation. After the unchanged program executes, a replacement overwrites its integral or inverse output with the corresponding reference calculation from graph-decoded preparation and task inputs. The other original output is preserved; recomputing preparation can change both. Multiple violations lead to abstention; a domain or execution error gives no measurable diagnostic; no detected violation does not certify correctness. Reference endpoints, case labels and post-repair outcomes are excluded from the selector's inputs. This exclusion does not make the retrospectively selected tasks a prospective evaluation.

Table 8: **Offline tool interventions on two archived programs after graph encoding.** Output pairs are (integrated volume, target time) for flow and (integrated response, target coordinate) for calibration, in each task's declared units. Independent task calculations check both outputs. Both selected repairs are correct; no change, a deterministic retry, an identity intervention and replacement of the other component each leave both cases incorrect. Replacing both components also repairs both cases. These are two retrospective examples, not a population success rate or evidence of optimal repair cost.

Archived program	Original (total, inverse)	Selected replacement	Repaired (total, inverse)
Flow	$(9, 25/6)$	Inverse calculation	$(9, 13/8)$
Calibration	$(15, 1/2)$	Integral calculation	$(45, 1/2)$

The complete roster for this offline illustration contains 11 cases: the two archived programs and nine constructed controls. It does not cover all upstream experiments. Its 421 offline endpoint invocations include 152 source-perturbation settings evaluated before and after the selected action at steps ± 0.1 and ± 0.05; these invocations are not independent tasks. Table 9 compares joint correctness of the total and inverse outputs in each case's specified input context, using the saved outcomes and

independent task calculations. The accompanying data include every case definition and outcome record. Independent grading compares both outputs with reference values calculated from each case's declared source data.

Table 9: **Joint-correct outcomes in the eleven-case offline illustration.** The last row identifies the unscored member of the nine controls; it is not an additional case. The action columns report jointly correct / scored cases; "not scored" denotes an unsupported input domain, not a failed endpoint. The constructed controls include two already-healthy cases, so outcome correctness is not a repair-success rate. The historical $0/18$ accuracy result belongs to the separate upstream development experiment.

Offline subset	**Scored / roster**	**Selected action**	**Replace both**
Archived baseline programs	$2/2$	$2/2$	$2/2$
Constructed controls	$8/9$	$4/8$	$6/8$
Unsupported-domain control (included above)	$0/1$	Not scored	Not scored

The nine constructed controls retain two already-healthy cases with no selected intervention, two repaired preparation errors, two missed defects, one incomplete repair with a second undetected defect, one ambiguous case and one unsupported-domain case. The last has no assigned score; it is not counted as an endpoint zero. Thus a passed task relation can miss an error, and correcting one detected violation need not repair the complete answer. The same relations and action rules can be evaluated by ordinary diagnostics given the same measurements; this executable illustration of diagnosis and intervention establishes no superiority or exclusive algorithmic advantage.

Transport controls and representation limits. Blocking preparation transport yields zero decoded preparation and triggers the domain guard before either archived program executes. This is an access and domain check, not numerical attenuation or a workability-boundary crossing. Blocking worksheet transport leaves these preparation-reading programs unchanged; separate constructed worksheet-reading consumers fail when worksheet transport is blocked. Neither intervention falls back to the original preparation or worksheet. Restoring the baseline graph restores the original, incorrect outputs. Transplanting the graph from a perturbed source into an unchanged raw-input context reproduces the corresponding deterministic source-rerun endpoint for these programs. These interventions establish the executed input dependency at this scope, not unrestricted mediation by the graph.

An archived pair with identical worksheets and program bytes but different preparation values gives different endpoints. The expanded graphs differ in the preparation coordinates, preserving the counterexample to a worksheet-only state description. A separate constructed program reads a raw sensor directly: holding the entire graph fixed while changing that sensor changes its total from 1.8 to 2.0. This is a counterexample to graph-only sufficiency for the wider permitted-input class, not a natural source rerun with preparation recomputed, nor a counterexample to the narrower preparation-only programs. The six independent encoded-input directions have zero state-response kernel, so they provide no informative positive sufficiency test. After replacing the flow inverse calculation, a source direction $(1, 0, -1)$ instead preserves total 9 while the inverse changes from $13/8$ to 1.65 or 1.60 at full steps (1.6375 or 1.6125 at half steps). This is a null direction for one endpoint, not for the retained state or the other endpoint.

The independent reproduction check reconstructs events, graph edges, both propagation ticks, state decoding, arithmetic, action selection and outcome grading without importing the production graph engine or selector. The result is an executable example of diagnosis and tool repair through the source–state–readout chain on these tasks. It neither revises the preceding AI measurements nor establishes fresh-model repair, unique fault attribution, general state sufficiency or live multi-agent dynamics.

E.2.2 Prospective four-case diagnostic pilot

The retrospective example leaves open whether its checks identify defects in programs generated after the rules are fixed. A separate pilot therefore uses four new numeric settings, two per existing

task family, fixed independently of observed model outputs. The three checks, action menu, cases and analysis code were committed before generation. Qwen2.5-14B generated one preparation and one program per case: eight fresh calls, without retries or replacement cases. All four decisions were recorded before intervention outcomes were evaluated. The four cases, rather than the 64 deterministic graph executions across intervention arms, form the case denominator. The settings are new; the task families and checks are unchanged.

Table 10 reports the complete panel. The graph and propagation rule are externally specified as in the preceding example. Interventions overwrite selected outputs with reference formulas applied to graph-decoded preparation; generated program code remains fixed. This is prospective selection of a formula-based intervention, rather than model self-repair. Both initially correct cases remain correct without intervention. The selector flags one of the two faulty programs and corrects one integral component, while both inverse defects are missed. It fully repairs $0/2$ faulty programs and changes jointly correct outcomes from $2/4$ to $2/4$. Correct components increase from $5/8$ to $6/8$. There are no interventions on the two healthy cases and no damage to the five initially correct components; these are descriptive counts on this panel.

Table 10: **Complete four-case prospective diagnostic panel.** Pairs are (integrated volume, target time) for flow and (integrated response, target coordinate) for calibration, in each task's declared units. Reference answers are reconstructed independently from the inputs using rational arithmetic. Every case is scorable; missed defects and incomplete repairs remain included.

Case	Baseline	Reference	After selection	Decision and outcome
Flow 01	$(44, 6)$	$(44, 13/6)$	$(44, 6)$	No action; inverse defect missed.
Flow 02	$(79/2, 20/3)$	$(79/2, 20/3)$	$(79/2, 20/3)$	No action; healthy case preserved.
Calibration 01	$(18, 6/13)$	$(18, 6/13)$	$(18, 6/13)$	No action; healthy case preserved.
Calibration 02	$(25/4, 4/3)$	$(95/4, 3/2)$	$(95/4, 4/3)$	Replace integral; inverse defect missed.

The other fixed arms retain the same four-case denominator. Unchanged execution, deterministic repeat, identity sham, and rebuilding preparation and worksheet each give $2/4$ jointly correct outcomes. Replacing every integral also gives $2/4$, replacing every inverse gives $3/4$, and replacing both outputs gives $4/4$. The last arm contains both correct task formulas; it is an independently checked arithmetic control, not an ordinary diagnostic policy. No separately implemented ordinary diagnosis was compared.

The detected violation illustrates the diagnostic role of a missing response. Calibration 02 has sensor means $(5, 12.5, 17.5)$; its generated integral is $(p(1) - p(0) + p(2) - p(1))/2$, where p reads the graph-decoded preparation. A uniform source shift of 0.1 changes the retained coordinates but cancels in that expression, giving zero integral change instead of the contract's 0.2 over an interval of length two. Integral replacement returns $95/4$, but leaves the wrong inverse $4/3$ unchanged. This value lies within $[0, 2]$, and Flow 01's wrong inverse 6 lies within $[0, 8.5]$: range checking misses both. Correct range membership therefore does not establish inverse correctness.

Whole-graph transplants with raw context fixed change at least one readout in each case. In Calibration 02 the diagnosed integral has zero response in the tested direction; the inverse changes instead. Blocking preparation transport makes the input inadmissible; blocking worksheet transport leaves this program unchanged. Restoring the graph restores its original wrong outputs. The fixed positive criterion requires both full repair and a resolved pre-intervention graph response of the diagnosed readout; no case satisfies that conjunction. Although zero response diagnoses the violated integral contract, the criterion is retained unchanged. These controls establish the implemented input route, with retained-state sufficiency and unique causal mediation unresolved. The accompanying code independently reconstructs decisions, graph executions and exact answers from all raw records and checks their hashes, without importing the original experiment implementation. The panel demonstrates a local correction and missed defects under fixed rules, without establishing reliable complete diagnosis across unknown tasks or a comparative advantage.

E.2.3 SELECTED-PATH MEASUREMENTS

Eight loading paths were recorded from file-grounded spreadsheet tasks in the typed-tool interface. Seven were selected retrospectively for comparison with the candidate monotone-response criterion: one locomotive spreadsheet read at five task-set widths and one path for each of two additional spreadsheets. One path records the longest observed sequence, exact–partially-exact–failed, and skips the workable-but-inexact boundary case; the other six begin in the partially exact region and record continuations to failure. No selected path traverses all four parts of the theoretical classification. The task-level readouts are the recorded tool-use indicator and its conjunction with answer correctness. Interior values of P_f, Q_f and r are pooled over framework–task records, rather than smooth changes of an individual run; tool use alone does not establish a successful tool execution. These runs provide support only for the applicable path segments. The eighth run, on a fourth spreadsheet, has readout $(w, r) : (1, 5/6) \to (1, 1) \to (0, -)$, which is non-monotone at the intermediate load point—r rises from $5/6$ to 1, after which $w = 0$ leaves r undefined. Thus seven of eight recorded paths are compatible with the candidate monotone-response assumption, while one is a counterexample to monotone readout response as the recorded tool-step budget varies for that task. Because the seven-path subset was selected retrospectively, the count is diagnostic rather than an independently timed test.

The seven selected paths cover three GAIA-derived, author-constructed spreadsheet task neighborhoods. The ordering-path analysis and replay of seven records from the live Qwen run use different evidence units, although four official task IDs contribute to both provenance chains. The live-origin records are used only for replay under fixed graph dynamics.

The ratio $r = Q_f/P_f$, defined where $P_f > 0$, uses pooled exact and workable framework–task records; representative partially exact occupancy occurs at $r = 11/18$. Per-instance traces are kept as measurement records and are not used in the stated theoretical comparison.

E.2.4 EXPLORATORY COMPARISON OF PERSISTENCE CONDITIONS

On one rule-gated eligibility-count task, carryover ablation, contradiction masking, mixture-of-agents and single-agent alternatives are compared with a persistent typed-state reference. The criterion requires recovery without greater resource use. For condition s, let R_s collect the coherence summary, contradiction handling and trajectory-shape indicator, and let L_s collect measured elapsed time and the counts of language-model calls, tool calls and feedback steps. Define

$$s \succeq_{\text{joint}} s' \quad \iff \quad R_s \geq R_{s'} \text{ componentwise and } L_s \leq L_{s'} \text{ componentwise,}$$

with at least one strict inequality for strict improvement; s' is the persistent reference. All coordinates enter the comparison, so matching the endpoint alone does not establish substitution. The source-reported trajectory-shape indicator has no public operational decision rule and cannot be independently reconstructed. Recorded execution fields include timeout, model-call budget, operator count, iteration limit, tool steps and token budget, but do not specify the complete execution conditions.

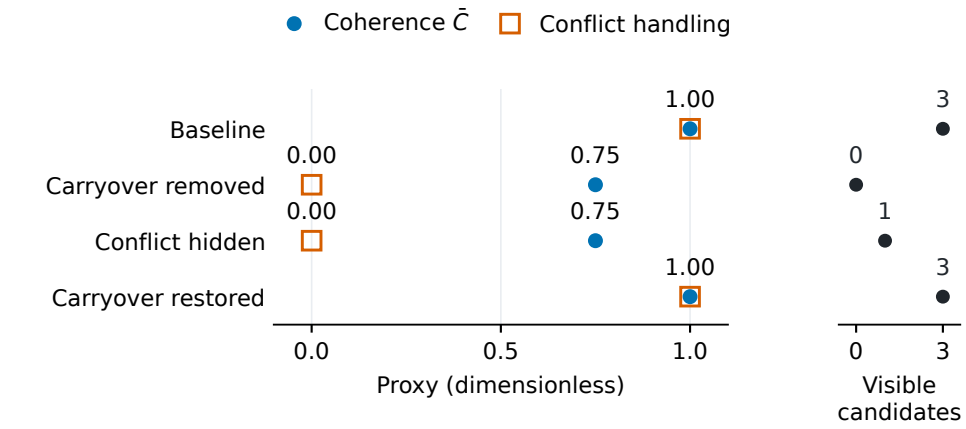
No displayed alternative satisfies the joint criterion (Fig. 4). The fuller source report and neighboring single-snapshot record are not released. These point estimates on one task therefore support neither a source-level reconstruction, a second matched comparison, a family-wide result, necessity nor a common ordering law.

E.3 CROSS-FAMILY COMPARISON AND PHYSICAL REALIZATION CHECKS

The following comparison asks which recorded AI outcomes occupy the same readout regimes and which paths meet the conditions for a boundary comparison. The physical realizations then provide separate dynamical and steady-state checks. Shared coordinates alone do not establish a common dynamical law.

E.3.1 REDUCED-COORDINATE COMPARISON

Planning and selected file-grounded records show how outcomes change along loading paths. The code-repair graph, constructed using endpoint information, provides an implementation check; de-



b **Comparison with persistent reference**

	Conflict hidden	Carryover removed	Mixture of agents	Single agent*
Coherence $\bar{C}$	×	×	×	×
Conflict handling	×	×	×	×
Trajectory shape	×	×	×	—
Elapsed time	×	×	×	✓
Model calls	✓	×	×	✓
Tool calls	✓	✓	✓	✓
Feedback steps	✓	×	×	✓
Joint criterion	×	×	×	×

✓ No worse × Worse — Unavailable

* Single-snapshot record

Figure 4: **Exploratory persistence interventions on one counting task.** **a**, Dimensionless coherence $\bar{C}$ (filled circles) and contradiction handling ("Conflict handling"; hollow squares), with candidate counts on the right. Coincident markers indicate equal proxy values. Conditions remove, hide or restore carryover. The released summary reports these proxy scores without operational scoring formulas or denominators; they are not success probabilities. $\bar{C}$ differs from the state-level functional $C(t)$ (Section B.3). **b**, Four alternatives compared with the persistent reference. Recovery rows test $\bar{C}$, contradiction handling and the source-reported trajectory-shape indicator; load rows compare elapsed time, language-model calls, tool calls and feedback steps. A check means recovery at least as high or load no greater; a cross means worse. The dash denotes unavailable trajectory shape for the single-snapshot record, not zero. The joint row requires every recovery and load condition; no alternative satisfies it. These one-task point estimates have no uncertainty intervals. Incomplete source records and an unspecified trajectory-scoring rule prevent full reconstruction of the comparison or inference that persistence is necessary across the task class.

layed counting records workable but inexact outcomes. These readout coordinates do not establish that the four families follow the same path to their first boundary or obey a common transport law.

The comparison distinguishes three operations. The state representation or readout can change at a fixed instance. A loading continuation varies the declared control while holding both fixed. Interventions on the medium include ablation, masking or restoration of persistent carryover. A coordinate change denotes only an invertible reparametrization; a change of representation or readout can discard information. Comparing a boundary step across families additionally requires matched readout definitions, loading semantics, and stopping conditions. The available records do not provide such overlap across the full family set.

For each family f, let b label its recorded loading setting, with $P_f^{\mathrm{rec}}(b)$ and $Q_f^{\mathrm{rec}}(b)$ the recorded readouts under that family's declared definitions. The recorded readout coordinates are

$$w_f(b) := P_f^{\mathrm{rec}}(b), \qquad r_f(b) := \frac{Q_f^{\mathrm{rec}}(b)}{P_f^{\mathrm{rec}}(b)} \quad (P_f^{\mathrm{rec}}(b) > 0).$$

This definition requires neither an equilibrium branch nor matched state coordinates across families. The loading label alone does not establish a monotone response. If a recorded sequence with fixed family semantics is specified before the crossing data are read, starts in the exact stratum, has no re-entry, and its recorded first-hit indices additionally satisfy $s_{EW} < s_{WR} < s_{RO}$, with all three transitions recorded, its regime order is:

$$\begin{aligned}&\text{exact stratum } (w > 0,\, r = 1) \;\rightarrow\; \text{partially exact region } (w > 0,\, 0 < r < 1)\\ &\rightarrow\; \text{workable-but-inexact stratum } (w > 0,\, r = 0) \;\rightarrow\; \text{failed stratum } (w = 0).\end{aligned}$$

Preparation P_f^{prep} initializes the execution; subsequent dynamics and the declared source history determine the augmented state s_T. The endpoint is specified on that state as $O_T(s_T)$, while $z_T = G_f^{\mathrm{cg}}(s_T)$ is the candidate retained coordinate. Where the endpoint factors through this coordinate, write $O_T = \bar{O}_T \circ G_f^{\mathrm{cg}}$. A declared nested pair $(P_f(z), Q_f(z))$ supplies the separate classification $(w, r) = (P_f, Q_f/P_f)$ on $P_f > 0$. Dependence of $\bar{O}_T$ on that pair alone is an additional realization-specific assumption. The readout level sets define the local boundaries, while the response map describes how perturbations move the retained state and its readout relative to them.

All four AI families permit coordinate placement; only explicit continuations support path statements. Matched adjacent-boundary steps and overlapping local regions across distinct mechanisms are not established.

E.3.2 Reported reduced coordinates and constructed controls

Table 11 reports family-specific (w, r) coordinates from planning, file-grounded and delayed-counting outputs, alongside the constructed code-repair control.

File-grounded fractions pool three frameworks over four tasks in the narrow neighborhood (12 records) or twelve tasks in the widened neighborhood (36 records). Tool-step budgets 2 and 1 give respectively 12 and 10 exact outcomes among 12 workable records in the narrow neighborhood. In the widened neighborhood, the corresponding counts are 25 and 22 among 36 workable records; thus $11/18$ denotes $22/36$, not an 18-record sample. At budget 0, no record is workable and r is undefined. Planning values are source-reported aggregate coordinates; their underlying counts and numerical control settings are not supplied in the accompanying evidence. Regime occupancy does not establish equilibrium branches or a common weakening sequence. Broader calibration requires denser boundary and response measurements, controlled adjacent-boundary continuations, overlapping matched regions and a rule relating those regions across realizations. Sampling density alone is insufficient. Such a comparison must measure the separately evaluated endpoint $O_T(s_T)$ alongside regime and boundary changes under matched conditions.

E.4 Physical realizations

The physical realizations remove the AI source-interface layer and serve different roles. In the driven delayed consensus network, the 64-site response is summarized by $(\mathcal{A}, \phi, V)$, with amplitude and

Table 11: **Representative readout regimes.** Dimensionless coordinates $w = P_f$ and $r = Q_f/P_f$ for nested outer workability and inner exactness readouts. The code-repair row is an implementation check constructed using endpoint information; the other rows summarize recorded outputs. Control settings are family-specific: reduced resources and increased stress are distinct interventions. The table reports representative readout regimes, without assuming a common equilibrium branch; r is undefined when $w = 0$, so the failed rows use an em dash.

Family and representative instance	**Control setting**	w	r	**Regime**
File-grounded typed-tool: file-grounded instance, narrow neighborhood	full	1.0	1.0	exact stratum
File-grounded typed-tool: file-grounded instance, narrow neighborhood	reduced	1.0	5/6	partially exact region
File-grounded typed-tool: same grounding record, widened neighborhood	reduced	1.0	11/18	partially exact region
Planning: constraint-planning neighborhood	full	1.0	≈0.67	partially exact region
Planning: constraint-planning neighborhood	thinned	1.0	≈0.44	partially exact region
Planning: constraint-planning neighborhood	minimal	0.0	—	failed stratum
Code repair (constructed check): verifier-bound repair neighborhood	thinned	1.0	0.0	workable-but-inexact stratum
Contradiction-rich delayed counting: contradiction-density neighborhood	stress	>0	0.0	workable-but-inexact stratum
File-grounded typed-tool: representative minimal-resource case	minimal	0.0	—	failed stratum

phase compared with their analytic consensus-mode predictions, while the delay history remains part of the retarded state. The external-field flock is a nonlinear steady-state example: its alignment fraction and polarization have separated sampled scales on $\eta \in [0.15, 0.70]$. These standard model classes are usually analyzed for synchronization, stability or collective order (Hale & Verduyn Lunel, 1993; Fridman, 2014; Niculescu, 2001; Olfati-Saber & Murray, 2004; Kuramoto, 1984; Pikovsky et al., 2001; Vicsek et al., 1995; Toner & Tu, 1995); the added measurement is the response of declared macroscopic readouts. The flock pair is not a nested measure and does not enter the four-regime classification; the two systems are not asserted to share a kernel, state coordinate or endpoint pair.

E.4.1 Driven delayed consensus network

Uniform driving of a symmetric delayed network provides an analytic calibration: does its spatial response concentrate in the consensus mode, and do finite-window amplitude and phase agree with harmonic balance? The driven mean is periodic, so this comparison uses a harmonic response rather than an equilibrium resolvent. The full state retains the coupling history; the response summary defined below is not assumed to close autonomously.

The realization is a driven delayed consensus network on a periodic degree-four ring graph with 64 sites,

$$\dot{c}_i(t) = -a\, c_i(t) + \frac{K}{|N_i|} \sum_{j \in N_i} c_j(t-\tau) + F\cos(\Omega t) + \sigma \xi_i(t), \tag{E.1}$$

with $a = 1.0$, $K = 0.40$, $F = 0.25$, $\Omega = 0.35$, and $\sigma = 0.015$ on $N = 64$ sites. The state $c_i(t)$ is a scalar activity at node i; N_i is the neighbor set of site i on the periodic degree-four ring, with cardinality $|N_i| = 4$, so the degree-four averaging operator $\frac{1}{|N_i|}\sum_{j\in N_i}(\cdot)$ (unit row sum), scaled by the coupling strength K, transfers delayed neighbor activity; $\tau > 0$ is the transport (communication) delay on the coupling, the swept variable of the validation; the drive supplies a coherent forcing at frequency Ω; and $\sigma\xi_i(t)$ is per-site white noise with $\langle \xi_i(t)\rangle = 0$ and $\langle \xi_i(t)\,\xi_j(t')\rangle = \delta_{ij}\,\delta(t-t')$ (δ_{ij} the Kronecker delta on the N sites, $\delta(t-t')$ the Dirac delta in time), i.e. independent, zero-mean, unit-intensity noise terms. The delayed history $c(t-\tau)$ is evaluated by linear interpolation in the stored trajectory, approximating the delayed argument of the continuous-delay model.

Observables, coarse-graining, and the retained state. The primary delay sweep measures the complex lock-in response $\widehat{Z}$ and compares it with the harmonic target $F\chi(\Omega,\tau)$, both defined below. It leaves outer workability saturated; separate drive and noise sweeps test nested readouts. The full simulated state contains the 64-site field and its delayed history. The declared readout first projects the field trajectory to the spatial mean

$$m(t) = \frac{1}{N}\sum_i c_i(t),$$

and then to the lock-in response $\hat{Z}$ at the drive frequency over a readout window of length T_{win},

$$\hat{Z} = \frac{2}{T_{\mathrm{win}}}\int_{\mathrm{window}} m(t)\, e^{-i\Omega t}\,\mathrm{d}t.$$

The 13-delay validation spans $\tau \in [0.43, 0.46]$ and uses the window $[300, 400]$, with $\Omega = 0.35$ and $\Delta t = 0.02$. Its numerical estimator is $(2\Delta t/T_{\mathrm{win}})\sum_{n=15000}^{20000} m(n\Delta t)e^{-i\Omega n\Delta t}$, including 5,001 samples. Amplitude and phase lag are averaged over 100 seeds at each delay before comparison with the harmonic target. The finite-window response is summarized by the three quantities

$$\boldsymbol{y}_{\mathrm{cons}} = (\mathcal{A}, \phi, V), \qquad \mathcal{A} = |\hat{Z}|, \qquad \phi = |\arg \hat{Z}|,$$
$$V = \frac{1}{NT_{\mathrm{win}}}\int_{\mathrm{window}} \sum_i \big(c_i(t) - m(t)\big)^2\,\mathrm{d}t.$$

The finite-window amplitude and phase estimates are compared with the harmonic targets $F|\chi(\Omega,\tau)|$ and $|\arg\chi(\Omega,\tau)|$; the window correction is derived below. These estimates are not a nested readout pair. The variance V measures the mean-square transverse fluctuations discarded by projection onto the spatial mean. Together, $\boldsymbol{y}_{\mathrm{cons}} = (\mathcal{A}, \phi, V)$ summarizes the effectively rank-one driven spatial response of the 64-site field. It is evaluated on the field trajectory over the full observation window, rather than on the terminal coupling history alone. Observation accumulators make it a terminal readout of the enlarged full state (Section B.2). It is not a three-dimensional autonomous Markov state: the full retarded state still includes the delay history $\zeta^t(\theta) = c(t+\theta)$, $\theta \in [-\tau, 0]$. This comparison requires neither an invariant slow manifold nor a time-summed resolvent.

Analytic consensus-mode prediction. For the spatially uniform mode the degree-four averaging has unit row sum, so the deterministic component (or ensemble mean) obeys the scalar delayed equation

$$\dot{m}(t) = -a\, m(t) + K\, m(t-\tau) + F\cos(\Omega t).$$

With the convention $m(t) = \mathrm{Re}\{F\chi(\Omega,\tau)e^{i\Omega t}\}$, harmonic balance gives

$$\chi(\Omega,\tau) = \big(a - Ke^{-i\Omega\tau} + i\Omega\big)^{-1},$$

so the deterministic consensus-mode amplitude and phase lag are

$$\mathcal{A}_{\mathrm{th}}(\tau) = F\left[(a - K\cos\Omega\tau)^2 + (\Omega + K\sin\Omega\tau)^2\right]^{-1/2},$$
$$\phi_{\mathrm{th}}(\tau) = \mathrm{atan2}\big(\Omega + K\sin\Omega\tau,\ a - K\cos\Omega\tau\big).$$

Routing of the noise. The model carries a per-site white noise $\sigma\xi_i(t)$ with the statistics stated above ($\langle\xi_i(t)\rangle = 0$, $\langle\xi_i(t)\,\xi_j(t')\rangle = \delta_{ij}\delta(t-t')$). The normal-mode (coupling-eigenvector) decomposition splits this noise into a zero-mode (consensus) projection and a transverse projection. The window summary $c|_{\text{window}} \mapsto \boldsymbol{y}_{\text{cons}}$ routes the zero-mode projection into the complex lock-in estimator $\hat{Z}$. Spatial averaging and the finite-window matched filter give an $\mathcal{O}(\sigma^2|\chi(\Omega,\tau)|^2/(N\,T_{\text{win}}))$ contribution to its variance, as derived below; this is not the amplitude or phase error scale. The transverse fluctuations are summarized by the window-averaged spatial variance V.

(i) Normal-mode decomposition. Write $c_i(t) = m(t) + \delta c_i(t)$ with $m = \frac{1}{N}\sum_i c_i$ and $\sum_i \delta c_i = 0$. The ring averaging matrix is symmetric with unit row and column sums, so it acts as the identity on the consensus mode and maps the zero-sum (transverse) subspace into itself; the drive $F\cos\Omega t$ is spatially uniform. Averaging Eq. (E.1) over i projects the dynamics onto the consensus mode and yields the scalar stochastic delay-differential (Langevin) equation

$$\dot{m}(t) = -a\,m(t) + K\,m(t-\tau) + F\cos(\Omega t) + \eta_m(t), \qquad \eta_m(t) \equiv \frac{\sigma}{N}\sum_{i=1}^{N}\xi_i(t), \tag{E.2}$$

the deterministic consensus-mode equation plus the zero-mode projection η_m of the noise.

(ii) Collective-noise intensity σ^2/N. Independence of the ξ_i gives

$$\langle\eta_m(t)\rangle = 0, \qquad \langle\eta_m(t)\,\eta_m(t')\rangle = \frac{\sigma^2}{N^2}\sum_{i,j}\langle\xi_i(t)\xi_j(t')\rangle = \frac{\sigma^2}{N}\,\delta(t-t'), \tag{E.3}$$

so the consensus mode sees white noise of collective intensity σ^2/N: the per-site amplitude σ enters the mean as an effective $\sigma/\sqrt{N}$, a factor $\sqrt{64} = 8$ here.

(iii) Lock-in as a finite-window matched-filter estimator. Equation (E.2) is linear, so $m = m_{\text{det}} + m_{\text{noise}}$. After transients, the periodic component is $m_{\text{det}}(t) = \text{Re}\{F\chi e^{i\Omega t}\}$, with $\chi = \chi(\Omega,\tau)$. On a rectangular window $[t_0, t_0 + T_{\text{win}}]$, its lock-in is

$$\widehat{Z}_{\text{det}} = F\chi + F\overline{\chi}\,e^{-2i\Omega(t_0+T_{\text{win}}/2)}\frac{\sin(\Omega T_{\text{win}})}{\Omega T_{\text{win}}}.$$

The second term is the contribution from the conjugate frequency of the real signal. It vanishes when $\Omega T_{\text{win}} \in \pi\mathbb{Z}$, and decays as the window grows. Thus $F\chi$, with amplitude $\mathcal{A}_{\text{th}} = F|\chi|$ and phase lag $\phi_{\text{th}} = |\arg\chi|$, is the harmonic-response target rather than the exact value of every finite window estimator. The recorded experiment uses the discrete rectangular sum

$$\widehat{Z}_{\text{grid}}(\tau) = \frac{2\Delta t}{T_{\text{win}}}\sum_{j=0}^{5000}\text{Re}\{F\chi(\Omega,\tau)e^{i\Omega t_j}\}e^{-i\Omega t_j}, \qquad t_j = 300 + j\Delta t, \quad \Delta t = 0.02.$$

Set $\mathcal{A}_{\text{grid}} = |\widehat{Z}_{\text{grid}}|$ and $\phi_{\text{grid}} = |\arg\widehat{Z}_{\text{grid}}|$. Across the 13 delays, $\phi_{\text{grid}} - \phi_{\text{th}} = 0.674400$–$0.676273°$, whereas the recorded seed-mean phase minus ϕ_{grid} is 0.058675–$0.058919°$. These signed terms add. For amplitude, the corresponding terms, both divided by $\mathcal{A}_{\text{th}}$, are -0.003809 to -0.003710 and $+0.001292$ to $+0.001299$; they partially cancel. These are analytic reanalyses of the stated estimator and stored means. The remainder combines effects such as discretization, residual transients and stochastic averaging; this decomposition does not identify their separate contributions. The stochastic part m_{noise} is the output of the stable filter $H(\omega) = (a - Ke^{-i\omega\tau} + i\omega)^{-1}$ (the continuation of χ to general ω) driven by white noise of intensity σ^2/N; by Wiener–Khinchin its power spectral density is the smooth bounded $S_m(\omega) = (\sigma^2/N)|H(\omega)|^2$. The lock-in integral is a Fourier estimator with bandwidth $\sim 1/T_{\text{win}}$.

For stationary noise with integrable autocovariance $R_m(v) = (2\pi)^{-1}\int S_m(\omega)e^{i\omega v}\,d\omega$, the complex variance means $\text{Var}_{\mathbb{C}}\,Z = \mathbb{E}|Z - \mathbb{E}Z|^2$. The rectangular-window integral has variance $4T_{\text{win}}^{-2}\int_{-T_{\text{win}}}^{T_{\text{win}}}(T_{\text{win}} - |v|)R_m(v)e^{-i\Omega v}\,dv$, hence

$$\text{Var}_{\mathbb{C}}(\hat{Z}_{\text{noise}}) = \frac{4\sigma^2|\chi(\Omega,\tau)|^2}{NT_{\text{win}}} + o(T_{\text{win}}^{-1}). \tag{E.4}$$

This asymptotic expression concerns the continuous rectangular-window estimator; the recorded discrete sum has its corresponding discrete covariance sum. Spatial averaging and narrowband estimation suppress the stochastic contribution by factors $1/N$ and $1/T_{\text{win}}$. The comparison with the

harmonic-response target also includes deterministic demodulation leakage, residual transients and numerical discretization. The reported relative amplitude error 2.51×10^{-3} and phase error 0.74° are the maximum discrepancies of the seed-averaged readouts over the 13 delays, and quantify the combined effect.

(iv) Transverse modes as the retained variance V. Subtracting Eq. (E.2) from Eq. (E.1) removes the uniform drive and consensus-mode dynamics exactly, leaving the transverse fluctuations forced by the transverse projection of the noise,

$$\dot{\delta c}_i(t) = -a\,\delta c_i(t) + \frac{K}{|N_i|}\sum_{j\in N_i}\delta c_j(t-\tau) + \sigma\,\delta\xi_i(t), \qquad \delta\xi_i \equiv \xi_i - \frac{1}{N}\sum_k \xi_k. \tag{E.5}$$

On the transverse coupling eigenmodes $\{e_q\}_{q\neq 0}$ (eigenvalues $\kappa_q \in [-1, 1)$; the consensus mode $\kappa_0 = 1$ is excluded), each amplitude δc_q is a stable scalar delay Ornstein–Uhlenbeck process whose stationary variance is the spectral integral

$$\langle \delta c_q^2 \rangle = \frac{\sigma^2}{2\pi}\int_{-\infty}^{\infty} \left|a - K\kappa_q e^{-i\omega\tau} + i\omega\right|^{-2} \mathrm{d}\omega =: \sigma^2\, g_q(\tau), \qquad g_q(\tau) > 0, \tag{E.6}$$

where $g_q(0) = 1/(2a_{\mathrm{eff,q}})$ is the memoryless Ornstein–Uhlenbeck factor and $a_{\mathrm{eff,q}} = a - K\kappa_q > 0$ for the stated parameters. By Parseval, the stationary expectation of the spatial variance is

$$\langle V(t)\rangle = \frac{1}{N}\sum_{i=1}^{N}\langle (c_i - m)^2\rangle = \frac{1}{N}\sum_{q\neq 0}\langle \delta c_q^2\rangle = \frac{\sigma^2}{N}\sum_{q\neq 0} g_q(\tau) = \mathcal{O}(\sigma^2). \tag{E.7}$$

The reported V estimates this stationary expectation by averaging the instantaneous spatial variance over the finite observation window. The transverse projection of the noise is the physical content of the third summary component V. The decomposition partitions the white noise into a coherent zero-mode contribution to $\hat{Z}$, admitted at order $\sigma^2|\chi(\Omega,\tau)|^2/(N\,T_{\mathrm{win}})$ in variance (Eq. (E.4)) and entering $(\mathcal{A}, \phi)$ as a sub-leading correction, and a transverse contribution with stationary variance $\langle V(t)\rangle = \mathcal{O}(\sigma^2)$. The analytic amplitude is $\mathcal{A}_{\mathrm{th}} = F|\chi(\Omega,\tau)|$. The stationary variance relative to its square is

$$\frac{\langle V(t)\rangle}{\mathcal{A}_{\mathrm{th}}^2} = \overline{g}(\tau)\,\frac{\sigma^2}{F^2|\chi(\Omega,\tau)|^2}, \qquad \overline{g}(\tau) = \frac{1}{N}\sum_{q\neq 0} g_q(\tau) = \mathcal{O}(1), \tag{E.8}$$

the mode-averaged factor of Eq. (E.7). The drive/noise ratio $\sigma/F = 0.06$ sets the prefactor $\sigma^2/F^2 \approx 3.6\times 10^{-3}$, and the measured value across the delay window is $V/\mathcal{A}^2 < 10^{-3}$, i.e. $\mathcal{A}^2/(\mathcal{A}^2+V) \geq 0.999$.

Reduction accuracy. Across the sampled delays, $\boldsymbol{y}_{\mathrm{cons}} = (\mathcal{A}, \phi, V)$ summarizes the stochastic 64-site network response; its amplitude and phase estimates are compared with the analytic consensus-mode predictions at matched τ. The maximum relative error in the amplitude $\mathcal{A}$ is 2.51×10^{-3}, the maximum absolute error in the phase lag ϕ is 0.74°, and the declared amplitude/variance diagnostic is

$$\frac{\mathcal{A}^2}{\mathcal{A}^2 + V} \;\geq\; 0.999$$

throughout the window, with $V/\mathcal{A}^2 < 10^{-3}$. This compares a time-averaged variance with the squared peak lock-in amplitude; it is not a physical time-averaged energy fraction. A pure sinusoidal mean has period-averaged square $\mathcal{A}^2/2$, so an energy interpretation would require a different weighting. A participation diagnostic follows from nonnegative weights

$$E_0 = \mathcal{A}^2, \qquad \sum_{k=1}^{N-1} E_k = V, \qquad r_{\mathrm{eff}} = \frac{(\sum_{k\geq 0} E_k)^2}{\sum_{k\geq 0} E_k^2}.$$

Writing $C = \mathcal{A}^2/(\mathcal{A}^2 + V)$, concentration of all transverse weight in one mode and equal distribution over $N-1$ modes give

$$\frac{1}{C^2 + (1-C)^2} \leq r_{\mathrm{eff}} \leq \frac{1}{C^2 + (1-C)^2/(N-1)}.$$

At the least-concentrated delay, $C = 0.9990030439$ and $N = 64$ give $1.0019959000 \leq r_{\text{eff}} \leq 1.0019968821$; both bounds round to 1.002. These bounds are derived from recorded amplitude and variance, not measured covariance spectra at all 13 delays. Their lower endpoint is not a lower bound for every delay. To express these same weights through eigenvalues μ_k of $N_{\text{samp}}^{-1} \sum_j \delta c(t_j) \delta c(t_j)^\top$, use $E_k = \alpha \mu_k / N$, where $\alpha = N_{\text{samp}} \Delta t / T_{\text{win}} = 5001/5000$. This factor matches the rectangular-quadrature definition of V; for seed-averaged records, average the covariance matrices over the same seeds. Unscaled covariance weights have a different normalization. The uniform drive excites only the consensus spatial mode. Its dominance is a spatial-response summary, not the minimal dimension of an autonomous three-dimensional state.

Activation of the endpoint readouts along two controls. The delay sweep above compares threshold-free amplitude and phase readouts with analytic predictions, while leaving workability saturated. A separate controlled measurement at $\tau = 0.43$ uses the readout thresholds fixed in the original delayed-consensus design:

$$P_f = \mathbf{1}[\text{stable},\ \mathcal{A} \geq 0.16,\ V \leq 0.05]\,,$$
$$Q_f = P_f\, \mathbf{1}[\mathcal{A} \geq 0.22,\ \phi \leq 35^\circ]\,,$$

and varies two physically different controls. The drive-amplitude sweep reduces the coherent signal at fixed noise $\sigma = 0.015$; the noise-amplitude sweep raises the transverse variance and perturbs the phase at fixed drive $F = 0.25$. Each control value uses the same 50 seeds. These are controlled retrospective measurements with the declared thresholds held fixed across both sweeps, not a prospective ordering test.

Averaging the binary readouts over the 50 seeds gives their support fractions $\widehat{P}_f$ and $\widehat{Q}_f$. Under decreasing drive amplitude, linear interpolation of these fractions gives half-support point estimates (where the fraction equals one half): $F \approx 0.163$ for Q_f and $F \approx 0.118$ for P_f. Under increasing noise amplitude, the corresponding point estimates are $\sigma \approx 0.138$ and 0.310. The representative support fractions around the crossings are listed in Table 12. Across 26 control values and 50 seeds per value, 442 of the 1,300 trajectories are workable but inexact, $Q_f = 0 < P_f$, and the nested implementation has no $Q_f > P_f$ violations. The half-support order is the same on both loading directions. Nesting guarantees the corresponding weak order; the controlled result measures a separation of the interpolated sample curves and the frequency of workable-but-inexact outcomes on these two declared paths. On the drive axis, the locations mainly reflect the amplitude thresholds. On the noise axis, 179 of 650 trajectories have $P_f = 1, Q_f = 0$: all violate the phase criterion and none violates the extra amplitude criterion. Another 150 have $P_f = 0$, which forces $Q_f = 0$ regardless of exactness; 68 of these have phase at or below 35°. Thus phase failure cannot explain all 329 cases with $Q_f = 0$. The outer failures exceed the variance threshold. The 20,000 paired-seed bootstrap resamples quantify crossing-location uncertainty rather than independently testing the weak order. The 95% percentile interval for the drive-axis exactness point is 0.1625–0.1629; the corresponding noise-axis interval is broad, 0.075–0.303. The reported half-support locations also depend on the estimator and phase criterion: inserting the noise-free harmonic solution into the stated 5,001-sample lock-in sum at $\tau = 0.43$ gives 34.819°, only 0.181° below the 35° exactness threshold. This small deterministic margin may affect the noise-axis crossing; attributing its broad interval to that margin would require a separate sensitivity comparison. The half-support locations are therefore descriptive interpolation estimates rather than precisely resolved critical thresholds. The bootstrap conditions on the control grid and observation window; it does not resolve threshold locations within an interval where all sampled outer outcomes change together.

Relation to standard consensus analysis. Characteristic-root and Lyapunov–Krasovskii methods analyze stability, while harmonic balance supplies the mean response (Hale & Verduyn Lunel, 1993; Fridman, 2014; Niculescu, 2001; Olfati-Saber & Murray, 2004). For the stated normalized coupling modes, $a > |K|$ gives delay-independent stability. The additional measurement here summarizes the driven spatial response by $(\mathcal{A}, \phi, V)$, checking amplitude and phase against harmonic balance and reporting transverse variance separately. These observables are also accessible to conventional consensus analysis; the calibration does not establish a new stability result or autonomous three-dimensional state.

Table 12: **Two distinct controls activate both consensus readouts.** Read the drive sweep from high to low F at $\sigma = 0.015$, and the noise sweep from low to high σ at $F = 0.25$; both use delay $\tau = 0.43$. $\widehat{P}_f$ and $\widehat{Q}_f$ are workable and exact fractions over the same 50 seeds per control point; $\langle \mathcal{A} \rangle$ and $\langle V \rangle$ are seed means of lock-in amplitude and transverse variance, not fractions. Values use the simulation's field and time units: F has field/time units, σ field/$\sqrt{\text{time}}$, $\mathcal{A}$ field units and V squared field units. The full record contains 13 points on each axis.

Control	**Value**	$\widehat{P}_f$	$\widehat{Q}_f$	$\langle \mathcal{A} \rangle$	$\langle V \rangle$
drive F	0.165	1.00	0.94	0.2254	1.15×10^{-4}
drive F	0.160	1.00	0.00	0.2186	1.15×10^{-4}
drive F	0.120	1.00	0.00	0.1640	1.15×10^{-4}
drive F	0.115	0.00	0.00	0.1571	1.15×10^{-4}
noise σ	0.100	1.00	0.56	0.3420	5.12×10^{-3}
noise σ	0.150	1.00	0.48	0.3422	1.15×10^{-2}
noise σ	0.300	1.00	0.46	0.3430	4.61×10^{-2}
noise σ	0.320	0.00	0.00	0.3431	5.25×10^{-2}

E.4.2 External-field Vicsek: separated steady-state readout scales and a normalization control

The preceding consensus realization measures a retarded finite-window response on a delayed state. The flock experiment measures steady-state polarization and field alignment in a polar flock weakly pinned by an external field. Its headings interact through a nonlinear circular-mean (Arg) update. On the reported interval, the alignment fraction first falls below one half at a sampled noise level where the polarization remains above one half. The accompanying normalized-profile comparison is made over shrinking neighborhoods of the loading coordinate, not over observation-time windows, and the imposed local linear form serves only as a normalization control.

Model and observables. For $N = 256$ self-propelled agents in a periodic box of side $L = 8$, with interaction radius $r_{\text{int}} = 1$, speed $v_0 = 0.3$, and a weak fixed aligning field of strength $h = 0.6$ at orientation $\theta_{\hat{e}} = 0$, the heading update is

$$\theta_i(t+1) = \operatorname{Arg}\Big(\sum_{j \in N_i} e^{\mathrm{i}\theta_j(t)} + h\, e^{\mathrm{i}\theta_{\hat{e}}} \Big) + \eta\, \xi_i(t), \qquad \xi_i \sim \mathcal{U}(-\pi, \pi), \tag{E.9}$$

with positions advected at constant speed v_0 and N_i the neighbors within r_{int}; the angular noise $\xi_i \sim \mathcal{U}(-\pi, \pi)$ is i.i.d. across agents and steps, with variance $\operatorname{Var}(\xi_i) = (2\pi)^2/12 = \pi^2/3$. Write $\delta_i = \operatorname{Arg} e^{i(\theta_i - \theta_{\hat{e}})} \in (-\pi, \pi]$ for the wrapped heading deviation from the field direction. The single loading parameter b is the angular noise η; the field strength h is weak and held fixed, so the loading knob is one-dimensional. Define the instantaneous polarization magnitude and aligned population fraction by

$$p(t) = \left| \frac{1}{N} \sum_{i=1}^{N} e^{\mathrm{i}\theta_i(t)} \right|, \qquad q(t) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[|\delta_i(t)| \le \text{tol}]\,.$$

The reported macroscopic readouts are

$$P_f = \langle \langle p(t) \rangle_{\text{time}} \rangle_{\text{seed}}\,, \qquad Q_f = \langle \langle q(t) \rangle_{\text{time}} \rangle_{\text{seed}}\,, \quad \text{tol} = 30^\circ = \tfrac{\pi}{6}. \tag{E.10}$$

The records use 500 burn-in steps followed by 300 measurement steps for each of four seeds. Neighborhoods include the focal particle, and positions advance with the updated headings. No observation-time or fixed-density size extrapolation is inferred from these records. The inner average is over recorded steps after transients within one simulation; the outer average is over seeds. Thus the particle mean is taken first, its modulus is evaluated at each step, and only then are time and seed averages formed. The modulus is not moved outside either average. The polarization P_f measures coherent order and Q_f measures alignment to the field within the angular tolerance (varied over $\{20^\circ, 30^\circ, 45^\circ\}$ in the threshold-robustness check), with dimensionless ratio $r = Q_f/P_f$ for $P_f > 0$. On the sampled grid, $Q_f \le P_f$ for $\eta \in [0.15, 0.70]$. The pair $(P_f, r) = (P_f,\, Q_f/P_f)$

separates the polarization amplitude from the aligned population fraction normalized by that amplitude. Because P_f is an order magnitude and Q_f a population fraction, r is the ratio of two distinct reductions of the same headings rather than a conditional exactness fraction; their ordering on this interval is empirical. As the loading η rises, the sampled interval exhibits separated losses of field alignment and coherent order. Outside $\eta \in [0.15, 0.70]$ the numerical order $Q_f \leq P_f$ fails: near disorder Q_f approaches the 30° random-orientation baseline $1/6$ while P_f falls toward a finite-size residual polarization, and at the lowest nonzero noise points Q_f can also exceed P_f. The flock pair is therefore not used as a nested workability–exactness coordinate. The circular-mean (Arg) coupling in Eq. (E.9) is nonlinear in the headings, so this is a genuinely nonlinear many-body realization of locally ordered readouts on the calibrated interval. The alignment fraction predicted by the one-parameter wrapped-normal summary inferred from P_f is compared with the measured Q_f below. This is a consistency description of two steady-state readouts, not a reduction to a dynamical retained coordinate: no dynamical soft mode, slaving relation or invariant slow branch is measured here.

Separated sampled readout scales. As η increases the alignment erodes before the polarization collapses. Over $\eta \in [0, 0.85]$ the polarization falls from $P_f = 1.000$ to 0.075 and the alignment fraction from $Q_f = 1.000$ to 0.178; with seeds 21–24, the alignment fraction first falls below one half at the sampled value $\eta = 0.35$, while the polarization first does so at $\eta = 0.60$, resolving two separated half-value scales on the sampled interval $\eta \in [0.15, 0.70]$ (Fig. 6c). On that interval the pair satisfies $Q_f \leq P_f$, but the ratio $r = Q_f/P_f$ is not monotone and is not used to infer a global loading law or four-part boundary regime. At angular tolerances 20°, 30° and 45°, the alignment fraction first falls below one half at sampled noise levels 0.25, 0.35 and 0.50, respectively. Each lies below the sampled noise level 0.60 at which polarization first falls below one half; all four seeds give these same sampled locations. At 45°, the two first-sampled half-value locations differ by 0.10, or two noise-grid steps. The periodic wrapped-normal summary below places equal half-value points at an angular tolerance of 45.50°, so the tested 45° readout lies close to the point at which the ordering can reverse. The result establishes separation only for the tested thresholds, not a threshold-independent ordering law. The observed separation is consistent with amplitude–direction separation in a pinned polar flock: local coupling maintains substantial collective order while alignment to the external direction weakens first. No dynamical eigenmode is inferred from this steady-state sweep.

Wrapped-normal consistency summary. For a wrapped-normal heading distribution centered on the imposed field, the population resultant is $R = |\mathbb{E}e^{i\theta}| = e^{-\sigma_\theta^2/2}$. The measured P_f instead averages the instantaneous finite-flock modulus; it is not generally R. Setting $R = P_f$ defines an effective summary, with inferred width (for $0 < P_f < 1$)

$$\sigma_\theta^{(P)}(\eta) = \sqrt{-2 \log P_f(\eta)}.$$

The corresponding probability for the principal wrapped angle to lie within the tolerance is

$$Q_f^{\mathrm{WN}}(\eta) = \sum_{k \in \mathbb{Z}} \left[\Phi_{\mathrm{N}}\left(\frac{\mathrm{tol} + 2\pi k}{\sigma_\theta^{(P)}(\eta)} \right) - \Phi_{\mathrm{N}}\left(\frac{-\mathrm{tol} + 2\pi k}{\sigma_\theta^{(P)}(\eta)} \right) \right], \qquad \mathrm{tol} = 30^\circ.$$

Here Φ_{N} is the standard-normal cumulative distribution function. Using the measured P_f values to infer $\sigma_\theta^{(P)}$, with no fit parameter, gives $\mathrm{RMSE}(Q_f, Q_f^{\mathrm{WN}}) = 0.0507$ on $\eta \in [0.15, 0.70]$. The residual includes finite-size effects, collective-direction fluctuations and distributional mismatch: a perfectly coherent flock can turn away from the field while its polarization stays one. This is a descriptive comparison of the two curves, not a distributional fit test, a measured dynamical state or a fluctuation–relaxation law.

Normalization control for readout profiles. This comparison measures the residual differences between two profiles after aligning their half-value locations and scaling by their selected bracketing secants. Its AI input is a deterministic replay of recorded model-prepared source events across endpoint acceptance thresholds. The graph scores remain fixed along this sweep; the threshold changes the selected fields. Here Q_f is the mean fraction of correctly decoded fields, distinct from whole-instance exactness. The flock input is the recorded steady-state noise sweep. Its readouts

satisfy $Q_f \leq P_f$ on the sampled interval $\eta \in [0.15, 0.70]$, but wider normalization windows also use records below $\eta = 0.15$. The profile comparison requires neither nested nor numerically ordered readouts. Let b be the AI acceptance threshold or the flock angular noise, $b_0 = 0$ the baseline control value, and $b_{1/2,f}$ the half-value location defined by $Q_f(b_{1/2,f}) = Q_f(b_0)/2$. For a differentiable decreasing profile with nonzero slope at this anchor, define the normalized support S_f, loading coordinate λ_f and response length ℓ_f by

$$S_f(b) = \frac{Q_f(b)}{Q_f(b_0)}, \qquad \lambda_f = \frac{b_{1/2,f} - b}{\ell_f}, \qquad \ell_f = \frac{Q_f(b_0)}{\left|\mathrm{d}Q_f/\mathrm{d}b\right|_{b=b_{1/2,f}}}, \tag{E.11}$$

The normalization fixes $S_f(b_{1/2,f}) = \frac{1}{2}$ and $S_f = \frac{1}{2} + \lambda_f + o(\lambda_f)$: each differentiable profile has the same value and unit positive slope in its own normalized loading coordinate at the half-value point. For the recorded profiles, the half-value anchor is linearly interpolated within the first adjacent pair that brackets the half-value, and that pair's secant replaces the derivative in ℓ_f. This aligns the anchors and selected secant slopes. The interpolated AI profile has different one-sided slopes at its anchor, so the reference line $S = \frac{1}{2} + \lambda$ is not a shared two-sided tangent. Table 13 and Fig. 5 compare residuals to this reference line and between the two profiles. The residuals between profiles use piecewise-linear interpolation in the dimensionless loading interval $|\lambda| \leq W$, where W is its half-width, on the closed symmetric grid $\lambda = k/40$, $|k| \leq 40W$, for integer k, with equal point weights. These residuals mix profile differences with sampling, interpolation and anchor estimation; they do not estimate universal curvature or response shape.

The finite-window dynamical checks belong to the separately reported graph and driven-consensus experiments. The threshold-replay curve here calibrates the readout, while the flock supplies steady-state curves. Their comparison shows what half-value and secant normalization remove and what residual variation remains; numerical equality of the profiles is not required by the response construction.

Table 13: **Loading-window dependence of the profile-normalization control.** Dimensionless half-width W selects $|\lambda_f| \leq W$; it is not observation time. Normalized support $S_f = Q_f/Q_f(b_0)$ is compared with $S_f = \frac{1}{2} + \lambda_f$, fixed by half-value and secant normalization. The three residual columns give dimensionless root-mean-square (RMS) differences. Recorded-point counts are given for flock and AI, respectively. Reference-line residuals use these points; flock–AI residuals use the equally weighted interpolated grid $\lambda = k/40$, $|k| \leq 40W$. Dashes mean fewer than two recorded AI loading levels, so that residual is undefined. At $W = 0.10$, only the two flock points defining the secant remain, so their reference-line residual is zero by construction. Residuals include sampling, interpolation and anchor effects; they do not estimate curvature.

Half-width W	**Points (flock / AI)**	**Flock–line RMS**	**Flock–AI RMS**	**AI–line RMS**
0.50	13 / 4	0.118	0.122	0.059
0.35	9 / 3	0.085	0.100	0.021
0.25	6 / 1	0.044	0.065	—
0.20	5 / 1	0.028	0.045	—
0.15	4 / 1	0.016	0.028	—
0.10	2 / 1	0.000	0.015	—

Relation to standard flock descriptions. Continuum descriptions retain density and vector polarization (Vicsek et al., 1995; Toner & Tu, 1995; Marchetti et al., 2013; Ramaswamy, 2010; Chaté, 2020). The scalar polarization magnitude alone gives equal values to aligned and misaligned polarized flocks. Reading the alignment fraction alongside it resolves that distinction on the same trajectories, with unchanged dynamics. The measured separation, wrapped-normal summary and normalization control characterize these sampled steady-state curves; they establish neither a critical point or exponent nor a dynamical slow mode, observation-time response or universal profile.

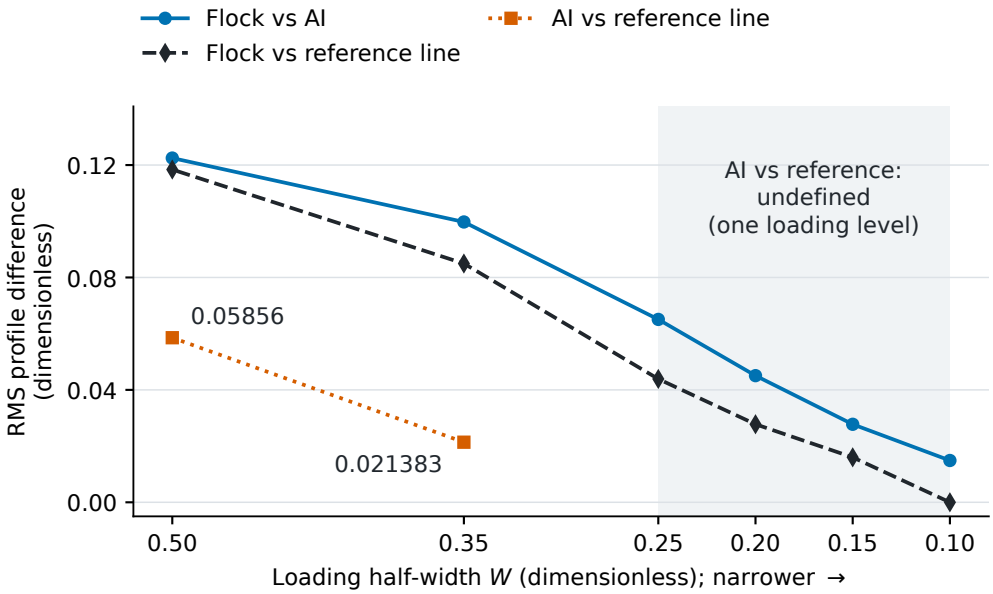


Figure 5: **Profile-normalization control.** Root-mean-square (RMS) differences between the flock and AI threshold-replay profiles, and between each profile and $S = \frac{1}{2} + \lambda$, as loading half-width W narrows. The full flock curves appear in Fig. 6c. Half-value alignment and bracketing-secant scaling fix the reference line, not a shared two-sided tangent. Table 13 specifies the interpolated comparison grid and recorded-point reference residuals. Shading marks $W \leq 0.25$, where the AI reference residual is undefined because only one AI loading level is recorded; it is not zero. All axes are dimensionless. Residuals mix profile, sampling, interpolation and anchor effects, rather than estimating curvature, common dynamics or a universal shape.

E.5 Finite contrasts and fresh-model intervention in a delayed-tool workflow

Table 14: **Operational requirements for response diagnosis.** These steps define what can be tested, rather than guaranteeing successful attribution for every workflow.

Required input	Controlled operation	Diagnostic output and limit
Source and operating conditions	Perturb a declared source; fix task, readout and permitted resources	Paired contrasts; insufficient resolution leaves the direction unresolved
Retained state and consumer inputs	Record formation and actual inputs; measure state and endpoint separately	Separates displacement from sensitivity; does not prove mediation
Candidate explanation and intervention access	Block, restore or replace an input; compare unchanged and alternative interventions	Tests that explanation; bypasses and indistinguishable alternatives limit attribution
Task specification and repair rule	Select before observing outcome; evaluate independently	Correct outcome, missed defect, partial repair or abstention; no violation is not a correctness certificate

This experiment measures changes in candidate states and model outputs under a source contrast, then evaluates whether a selected intervention improves the output. The apparatus comprises constructed inventory requests, virtual-time tool execution and fresh Qwen2.5-14B-Instruct-MLX-4bit

continuations. It measures model errors on synthetic tasks, not production incident frequencies. Four calibration cases represent irrelevant-warehouse returns, reversed arrival and revision order, a cancelled latest request, and repeated delivery. A separately authored roster and ordinary diagnostic rule were fixed with the response rule before model calls. Separate authorship within the AI-assisted study is not external replication or independent task custody.

System boundary and candidate states. Requests are issued before a snapshot is recorded. Each request carries an identifier, warehouse, revision and active or cancelled status. A fixed warehouse–revision lookup supplies its eventual quantity. The source contrast exchanges the warehouses of two active requests. Both query histories execute against the same lookup and produce identical numerical returns in the same arrival order. Thus the contrast changes a real query argument within the constructed environment, rather than attaching an incompatible history to a stored result. Delay is represented by virtual event order, not measured network latency. The later model reads the snapshot, complete request history and correctly identified returns.

The snapshot-only candidate z_P contains the target warehouse, requested quantity and current revision. Candidate z_H additionally contains the request records. Neither candidate restricts the model's input interface. An unrelated business note remains accessible to the model and is varied in a separate control. The source contrast leaves z_P unchanged and changes z_H. The readout is the reported inventory quantity and the allocate/hold action. The task rule selects the highest active revision for the target warehouse not exceeding the current revision; allocation requires sufficient quantity. Truth changes with the legitimate query contrast. The response is therefore measured in the reported quantity and action, separately from correctness scores.

Calibration criterion and outcome. Three fresh calls per source condition yield 24 calls. Temperature is 0.2, top-p is 1, the output cap is 128 tokens, the context is 8192 tokens, and calls are sequential with a 90-second deadline. Repetition measures observed variation without an assumption of independent random seeds. Baseline/probe order is shuffled using a fixed case-derived ordering. Raw output must be one exact-key JSON object; malformed outputs receive no semantic repair or silent retry.

All 24 outputs parse, but only 21 are jointly correct, below the fixed 22/24 instrument threshold. The scheduled four-case exploration and sixteen-case validation stages, totaling 320 further calls, are unrun. The criterion remains unmet. In cases 1, 2 and 4, all three measured quantity contrasts match the task contrast at the same snapshot. In case 3 the required contrast is $14 \to 9$, while all three measured pairs remain $9 \to 9$. The response rule selects a lossless request–return join for this zero response; the independently implemented ordinary rule makes the same selection by checking the raw request associations. Both select no change for the other three cases. These decisions precede all intervention outcomes.

Exploratory intervention follow-up. A separate plan, written after calibration failed, evaluates all four cases without changing their tasks, prompts or recorded selections. Its 72 fresh calls comprise three repetitions of six operations per case: an unrelated-note control, unchanged retry, lossless join, a readout checklist, both join and checklist, and idealized refresh. The join appends every request and its return as an associated row, preserving all records without selecting the correct row or supplying an answer. The checklist restates filtering, revision choice and threshold comparison. Refresh is an explicitly answer-informed comparison: a deterministic tool supplies the authoritative quantity. It is not an autonomous repair. Calibration and this follow-up comprise 96 model calls; no failed call is replaced. Each planned intervention is linked to the previously saved decisions.

The selected join corrects all three fresh continuations of the faulty case; unchanged retry and the checklist correct none. This supplies a concrete answer-free intervention result beyond reference-formula replacement. The ordinary rule selects the same operation, and applying the join to every case also succeeds. Consequently, the example establishes neither diagnostic superiority nor a need for adaptive selection. The complete deterministic task contract also admits a direct reference executor without any language-model call.

Costs and inference limits. The diagnostic measurements cost six calls per case. With illustrative utility $+1$ for a jointly correct result, -2 for unsafe allocation, -1 for invalid output and zero otherwise, subtracting 0.01 per deployment call and 0.05 per refresh gives mean utility

Table 15: **Calibration and exploratory intervention outcomes for all four cases.** The calibration row reports the baseline source condition. Calibration additionally includes twelve source-contrast calls; its full result is 21/24, below the fixed 22/24 threshold. Each entry counts jointly correct quantity and action among three repeated calls to one case; repeats are not independent task samples. The selected policy retains the nine correct calibration outputs in untouched cases and uses three fresh join outputs in case 3. Cases 1–4 concern irrelevant warehouses, arrival/revision order, a cancelled latest request and repeated delivery, respectively. Idealized refresh supplies the authoritative quantity and is answer-informed.

Operation	Case 1	Case 2	Case 3	Case 4	Total
Calibration baseline	3/3	3/3	0/3	3/3	9/12
Unrelated-note control	3/3	3/3	0/3	3/3	9/12
Fresh unchanged retry	3/3	3/3	0/3	3/3	9/12
Lossless join	3/3	3/3	3/3	3/3	12/12
Readout checklist	3/3	3/3	0/3	3/3	9/12
Join and checklist	3/3	3/3	3/3	3/3	12/12
Idealized refresh	3/3	3/3	3/3	3/3	12/12

$1 - 0.01(6 + 1/4) = 0.9375$ for either selected policy: all four cases incur six diagnostic calls, and one needs an additional intervention call. A fixed join or join-plus-checklist uses one call per case, giving $1 - 0.01 = 0.99$. These are equally weighted per-case policy costs, not the total cost of collecting all experimental arms. The weights are benchmark preferences, not measured economic or physical costs. No fixed strategy is selected on independent validation data. The three healthy cases are untouched by selection; their preservation is not an independent intervention result.

The calibration criterion was not met, and follow-up protocols were specified afterward; their results are exploratory, not independent validation. Repeated output contrasts at fixed snapshots demonstrate the usefulness of observing request history in this apparatus; the unrelated-note control does not certify z_H within a distributional equivalence margin. No first-order susceptibility, uniquely identified fault mechanism, positive state sufficiency, population repair rate or cross-model generalization is inferred. Reanalysis reconstructs the raw-call records, selection timing, full denominators and comparison outcomes from the anonymous supplement; its transformed export does not authenticate historical machine identity.

Paired response restoration on the known faulty case. A further exploratory test distinguishes correct answers at one input from recovery of the task-required source contrast. The case and useful join were known from the preceding calls. Before the new calls, the four operations, both source conditions, three repetitions, full randomized schedule and independent reconstruction were fixed. All 24 calls are fresh; historical outputs are excluded from these pairs. The source contrast, task, model and sampling settings are unchanged. Pair indices specify comparisons, not shared random seeds. The original calibration criterion remains unmet; the 320 later calls remain unrun.

The primary descriptive criterion requires all three join pairs to have both correct quantity/action endpoints and the exact signed quantity difference -5. Correct difference alone is insufficient: equal endpoint errors can cancel. Parser, source construction and record consistency are checked separately from consumer correctness. All 24 outputs parse; no call is replaced or excluded. Every join pair meets the criterion, whereas each control pair has difference zero (Table 16). Thus the lossless join restores a finite response to request–return associations already present in the reader's input. It precomputes those associations and may alter salience and prompt length; the result does not identify a unique internal mechanism.

Within each operation/source condition, the three recorded answer strings are identical. This describes repeatability within the run; identical outputs neither establish nor disprove sampling independence. All measurements concern one selected case. Every operation is already correct at the contrast endpoint $9/\mathrm{hold}$, so that endpoint alone does not distinguish the operations. The paired result excludes an always-14 response after joining, but this single warehouse exchange cannot distinguish the full task rule from a shortcut that uses the first returned record for the target warehouse. General use of revision, cancellation and cutoff rules remains untested by this contrast.

Table 16: **All paired follow-up outcomes on one previously examined case.** Entries are inventory quantity/action; each occurs in all three repetitions. Differences are source-contrast minus baseline quantity in inventory units. All four operations use three fresh pairs (24 calls total); pairs are not independent task samples. Truth is $14/\mathrm{allocate}$ at baseline and $9/\mathrm{hold}$ under the source contrast. A correct pair requires both reported endpoints, not just their difference.

Operation	Baseline	Source contrast	Differences	Correct pairs
Raw	9/hold	9/hold	0, 0, 0	0/3
Lossless join	14/allocate	9/hold	$-5, -5, -5$	3/3
Checklist	9/hold	9/hold	0, 0, 0	0/3
Unrelated note	9/hold	9/hold	0, 0, 0	0/3

All arms use six calls. Server-reported input/output token totals are 2094/60 for raw, 2724/69 for join, 2364/60 for checklist and 2076/60 for unrelated note: 9507 tokens in total. Summed client durations are respectively 2.885, 3.463, 3.666 and 3.787 seconds (13.800 seconds total). These are observed resource counts on one machine, not deployment cost estimates; cache and runtime variation preclude a latency ranking from these calls. Three repeated pairs describe this known case only. They establish neither distributional equivalence, a repair probability on new cases, nor an advantage over the ordinary selector or a fixed join.

E.6 Development study of response-guided action selection

This separate development study asks whether response measurements help select an intervention in scientific-claim verification. In SciFact (Wadden et al., 2020), fixed BM25 retrieval supplies six abstracts. A Qwen2.5-14B-Instruct model with 4-bit MLX weights selects sentence identifiers; a source-checked parser resolves them to ordered original text. A second call reads only the claim and this evidence, returning support, contradiction or insufficient evidence. Prespecified format conversions preserve source text; invalid selections remain failures, distinct from valid empty selections. After an invalid selection, the dependent reader is not called and its outcome is recorded as unavailable. Serial requests use fresh caches, explicit random seeds, temperature 0.7, no probability truncation or repetition penalties, and output limits of 512 writer and 16 reader tokens.

The five actions are to keep the original answer, reselect evidence upstream, change the readout prompt with evidence fixed, read all six abstracts directly, or aggregate the upstream and altered-readout answers. Aggregation returns their common label when both are valid and agree; otherwise a further call combines their visible evidence and answers without reference labels. Consistency voting instead uses up to eight readings of the original evidence. All action outcomes are scored on the original, undeleted task.

Ordinary features (B) include output validity and label, evidence lengths and selection by retrieval rank, retrieval scores, available model-reported confidence, missingness indicators, and the selection change computable from deletion alone. Two source probes remove the first-ranked abstract or the first three, without replacement. These are finite contrasts, not derivatives. Endpoint features (E_1, E_2) add label and model-reported score responses and repeat controls; state features (S_1, S_2) additionally measure reselection beyond mechanical deletion. The one-probe pair uses the first probe and a fixed-evidence rereading; the two-probe pair also uses the second probe and a repeated writer–reader execution. Their maximum diagnostic costs are five and nine calls, respectively. Consistency features (K) add repeated-answer statistics to B. Matched E/S families share all other values and availability indicators.

The roster selected one training-split claim per citation/duplicate-connected source group by fixed hash order, excluding pilot and earlier design-example groups. Execution stopped at 89 complete tasks, one partial task and six unrun tasks out of 96. A post-interruption amendment included the completed initial segment of the roster before scoring; incomplete and unrun tasks are neither scored nor imputed. The two action panels are within-task repeats, with some shared calls: there are 89 groups, not 178 independent tasks. For each task, the harm, no-change and repair targets are defined relative to its original answer and averaged over its two panels. Six ridge predictors use five fixed group folds, training-fold median imputation and standardization, and penalties

$\{0.1, 1, 10, 100, 1000\}$ selected by Brier loss. Choices maximize predicted repair minus harm with fixed ties. Six sensitivity comparators fit single-split probability rules, including a constant candidate, using training-only features, thresholds and leaf frequencies on the same folds. Ridge tuning reuses the assessment folds; the accuracy-first fixed action is also selected on development outcomes. Table 17 therefore reports descriptive development scores, not independent confirmation. The stored chronology does not independently verify that reference labels were withheld during execution or establish the provenance and integrity of the original requests.

Table 17: **All declared development comparisons.** Each panel has denominator 89; accuracy and costs average the two panels within each task, then the tasks (89 source groups, not 178 independent samples). Calls and input-plus-output tokens are mean costs from the start of a task, including selector measurements and the chosen action, with shared calls counted once. B denotes ordinary features; E_i adds endpoint responses; S_i also adds state-reselection responses; $i = 1, 2$ counts source probes. K adds consistency features. Grouped simple-rule pairs have identical scores and costs. Failed dependencies can reduce actual calls below the nominal diagnostic maximum. Tuning and fixed-action selection use these development data; differences do not establish superiority.

Strategy	Correct panels 1, 2	Accuracy (%)	Mean calls	Mean tokens
Fixed actions and consistency voting				
Keep original answer	33, 33	37.08	1.49	2,977.60
Readout prompt intervention	33, 33	37.08	1.49	2,992.92
Upstream reselection	27, 28	30.90	1.43	3,047.46
Aggregate	54, 51	58.99	3.71	6,434.95
Direct reading (selected fixed)	65, 68	74.72	1.00	2,675.29
Consistency vote	34, 34	38.20	4.96	4,093.51
Ridge-based selection				
Ridge: ordinary B	66, 68	75.28	2.55	5,305.96
Ridge: endpoint E_1	65, 68	74.72	4.54	7,820.43
Ridge: state S_1 (secondary)	66, 69	75.84	4.54	7,819.17
Ridge: endpoint E_2	65, 68	74.72	7.58	12,636.39
Ridge: state S_2 (primary)	65, 68	74.72	7.56	12,662.37
Ridge: consistency K	63, 66	72.47	5.93	6,451.22
Simple-rule selection				
Simple rule: B	64, 66	73.03	2.18	4,726.80
Simple rules: E_1, S_1	64, 66	73.03	4.16	7,322.87
Simple rules: E_2, S_2	64, 66	73.03	7.25	11,848.90
Simple rule: K	64, 66	73.03	5.64	5,842.71

The primary S_2 policy has identical correctness to direct reading and E_2 on every panel, despite some different actions. Its request-duration sum averages 69.13 s per task versus 11.57 s for direct reading; these are not deployment latencies. The secondary S_1 gain of 1.12 percentage points over direct reading and E_1 comes entirely from one task, correct on both panels under the readout intervention and incorrect under direct reading. It does not replace the primary comparison. Neither observed equality nor these development differences establish population equivalence or comparative superiority.

Initial writer outputs comprise 44 valid, 42 invalid and three capped/unavailable outputs. Post hoc classification of the 42 invalid outputs finds 24 sentence-object schema errors, eight failures to reproduce visible source text, six top-level key errors and four JSON parsing errors. Although the writer was asked for identifiers, these diagnoses describe its actual replies, including unsolicited sentence objects. Measurements of reselection beyond mechanical deletion are available on 23/89 and 28/89 tasks: 16 have both, 35 either and 54 neither. Across all operations in the complete cohort, 17 calls reached their output cap, including the three unavailable initial writers. All 89 tasks remain included; missing response measurements are not observed zeroes. Limited availability does not establish the cause of the absent primary benefit. Likewise, 37.08% versus 74.72% correctness for selected-evidence and direct reading compares execution paths; it is not a matched-state rejection of a representation.

The stopped run incurred 1,511 attempts, including one unknown completion. Known input and output token subtotals are 2,046,646 and 60,264; the unknown contribution is not zero. Analysis made no new model calls, and independent confirmation was not run. The anonymous evidence addendum supplies outcomes, policy choices and resource records for reconstruction; reconstruction does not constitute model re-execution or independent authentication of historical custody.

Scope of the empirical comparisons. Table 7 separates constructed calculations, replay, measured responses and diagnostic outcomes. Broader model-family robustness, additional source directions and independent cross-realization replication require new measurements. The profile normalization in Section E.4.2 aligns selected anchors and slopes; its residuals combine profile differences, sampling, interpolation and anchor estimation, rather than estimating curvature or a universal shape. Extending responses to a coarser or mesoscopic description additionally requires verified closure, stability and matched readout semantics. These conditions are not established across the reported realizations. Independent outcomes under fixed rules are needed to extend a diagnostic to new tasks (Section 6).

F Data and code availability

F.1 Data availability

The accompanying data and code package contains the derived arrays for the AI endpoint-composition check, replay of model-generated source records, and consensus spatial-response and two-axis readout tests. It includes the recorded protocol and provenance, complete 652-call ledger, source-seal ledger and analysis for the controlled live paired experiment, together with the flock readout curves and their wrapped-normal consistency check. Stored model calls permit independent reanalysis without rerunning the language model.

For the retrospective offline example, the package contains the archived programs, source inputs, graph states and complete diagnostic and intervention records. Its documentation identifies omitted upstream records and removed machine-specific metadata; scientific outcomes are retained. For the prospective pilot, it contains all four cases, the eight raw model request–response records, the fixed decision rules and all 64 graph executions, including the cases that fail the positive criterion. The package also includes the 96 fresh calls, complete controls and unchanged selections for the delayed-tool calibration and exploratory follow-up, plus 24 separately specified paired calls testing response recovery on the known faulty case. The calibration criterion was not met, and the validation stages were not run. For the evidence-selection development study, a separate anonymous component provides task and panel joins, recorded predictions and policy selections, request costs, and measurement-availability classifications. Its verification reconstructs the reported outcomes and costs; supplied classifications are distinguished from fields recomputed from model responses. These development records do not include an independently evaluated confirmation set. The package does not reproduce the broader AI benchmark families. No external archival DOI or data accession is asserted for materials outside this package.

F.2 Code availability

The accompanying verification code recomputes response-composition errors, structural-control failures, finite-graph kernels, live finite differences, source-seal timing, response counts, support fractions, nested-readout and ordering tests, sampled half-value locations and physical-response summaries from the supplied records. For the consensus tests, it reconstructs individual classifications and interpolated crossings; the bootstrap intervals remain stored resampling results and are not regenerated.

Separate checks evaluate the analytic diagnosis and repair illustration with exact arithmetic, independently reconstruct the offline diagnostic outcomes, and reexecute the bundled graph adapter. Verification of the four-case pilot checks the supplied records and their documented export hashes, and independently reconstructs all case results, including the failed positive criterion, from a relocated copy of the package. The anonymous review export replaces identifying metadata and updates dependent hash references; it is not byte-identical to the sealed original records. Numerical program code and scientific outcomes are unchanged. These checks use Python 3.10 or later and its stan-

dard library; the package documentation gives the execution instructions and identifies the records supporting each result.

The package also includes the original live paired experiment and analysis implementations, protocol specification, fixed selection rule, visible inputs and physically separate verifier inputs, with their recorded hashes. The broader research software and full figure-generation pipeline are not included and may be made available on reasonable request for scholarly verification.

G NOTATION AND TERMINOLOGY

The main text defines the symbols used in its argument at first use. This section collects core and realization-specific notation and the technical vocabulary for reference.

A typed-evidence field is an externally declared AI graph field whose events carry evidence modality and response role; it is neither a graph neural network nor a model hidden state. A loading parameter is a scalar or vector control varied about a fixed reference protocol. A finite-window response maps a source perturbation to retained-state and endpoint-readout changes during a fixed window.

A prescribed delayed input has a generator outside the modeled local operators; its time shift alone does not imply endogenous feedback. Retarded dynamics depend on internally generated past quantities whose required histories belong to the augmented state. Retarded response instead means that the response vanishes before the source acts, without implying endogenous delayed feedback.

An operating branch is a local regular branch of retained-state equilibria parametrized by control; stability is a separate condition. A slow manifold is a locally invariant manifold on which eliminated variables are slaved to retained coordinates, where such a reduction exists. A branch residual is the back-substitution error for a linear solve using assembled A, B; it does not check matrix assembly.

Table 18: **Selected symbols.** Full-state, retained-state and readout roles are distinguished; realization-specific meanings are identified explicitly.

Symbol	Meaning and scope
$c_t \in H$	shared coordination field in Hilbert field space H, the field component of the augmented state
$s_t \in X_f$	full augmented state containing the field, memories, required delay histories, direct readout inputs and observation records
$z_T = G_f^{\mathrm{cg}}(s_T) \in Z_f$	candidate retained state; may include the full field. Low dimension, endpoint sufficiency and autonomous closure each require separate justification
$O_T : X_f \to \mathbb{R}^d$	endpoint readout specified on the full augmented state; endpoint response is its change under a source perturbation
$\bar{O}_T : Z_f \to \mathbb{R}^d$	projected readout where $O_T = \bar{O}_T \circ G_f^{\mathrm{cg}}$ holds on the declared domain
$\mathrm{D}_s O_T$	endpoint projection; under differentiable factorization it equals $\mathrm{D}_z \bar{O}_T \, \mathrm{D}_s G_f^{\mathrm{cg}}$
$\bar{O}_{f_i,T}$, $\mathcal{Z}_i^{\mathrm{reach}}$	realization-specific projected readout and reached retained-state set in a two-realization comparison
$d \geq 1$, q	readout-component count and field-channel count; neither fixes augmented- or retained-state dimension
u_t, δu_t, Δu_t	controlled source coordinates, tangent perturbation and finite source contrast; u also parametrizes a fixed operating branch
$\mathcal{U}_T$	declared finite-dimensional source-perturbation space over the observation window
$\mathcal{K}_{z,T}, \mathcal{K}_{O,T}$	first-order maps from $\mathcal{U}_T$ to retained-state and endpoint displacements
$\mathcal{H}_T^{(1)}$	linear factor map from $\mathrm{Im}\,\mathcal{K}_{z,T}$ to $\mathbb{R}^d$, existing when $\ker \mathcal{K}_{z,T} \subseteq \ker \mathcal{K}_{O,T}$

Table 18: Selected symbols (continued).

Symbol	Meaning and scope
$\boldsymbol{y}_{\mathrm{cons}} = (\mathcal{A}, \phi, V)$	consensus window summary: amplitude, phase lag and transverse variance, evaluated on the observed field trajectory
$\sigma_{\dot{\theta}}^{(P)}$	flock width inferred from P_f; a steady-state consistency summary
r_{PR}	participation diagnostic using squared singular-value weights of the constructed response matrix
r_{eff}	consensus participation diagnostic using $E_0 = \mathcal{A}^2$ and transverse weights summing to V
$A_t, B_t, \Phi(T,t)$	linearized full-state transition, source Jacobian and ordered transition product
$A_t^{\mathrm{ret}}, B_t^{\mathrm{ret}}, \Phi_{\mathrm{ret}}(T,t)$	corresponding maps for a closed retained-state linearization, where that closure holds
$A, B, \rho(A)$	time-independent transition and source maps of the closed retained graph; spectral radius of A
L, J_i	intrinsic field operator and composite local source
R_i, π_i, W_i	local read, update and write maps
G_i	source-coupling operator for a modeled local operator
E_ℓ, τ_ℓ, y_ℓ	coupling, delay and prescribed signal for a delayed input whose generator remains outside the modeled local operators
ξ_a^t	history of an internally generated delayed quantity, retained in $s(t)$
$C(t)$	coherence functional, Section B.3
$F_{\mathrm{full}}(s; u, y^t)$	augmented controlled evolution, including memory laws and internal history transport; y^t is prescribed
$f_u(z)$	reduced drift in $\dot{z} = f_u(z)$, where the stated local reduction exists
P_f, Q_f (nested criteria)	outer workability and inner exactness; logical inclusion and normalization give $0 \le Q_f \le P_f \le 1$. Numerical order alone does not establish nesting
$\widehat{P}_f, \widehat{Q}_f$ (consensus)	sample support fractions obtained by averaging the binary readouts over the same 50 seeds
P_f, Q_f (flock)	time- and seed-averaged polarization and alignment fraction; these observables are not nested
$w = P_f$, $r = Q_f/P_f$	workability and inner-to-outer ratio for nested criteria; w remains defined at $P_f = 0$, where r is undefined
$r = Q_f/P_f$ (flock)	alignment fraction divided by polarization for $P_f > 0$; not conditional exactness or a canonical (w, r) coordinate
$P_f^{\mathrm{prep}}, G_f^{\mathrm{cg}}$	execution-initialization and full-to-retained-state coarse-graining maps; preparation is distinct from subsequent dynamics and scalar P_f
$\widehat{\chi}^G(t,t')$	finite-graph finite-window response
$\chi_*^G = (I - A)^{-1} B$	time-summed graph response for time-independent stable dynamics
χ_*	continuum stable-branch susceptibility under the stated conditions
$\Sigma_{EW}, \Sigma_{WR}, \Sigma_{RO}$	regular readout boundaries; the stated sequence requires actual crossings, no re-entry and ordered, separated thresholds on a fixed path
Θ_f, Λ_f, b_f	fixed instance neighborhood, interface and loading coordinate; together with the fixed readout definitions, these specify an admissible family

H PHYSICAL MEASUREMENTS AND READOUT-REGIME ILLUSTRATIONS

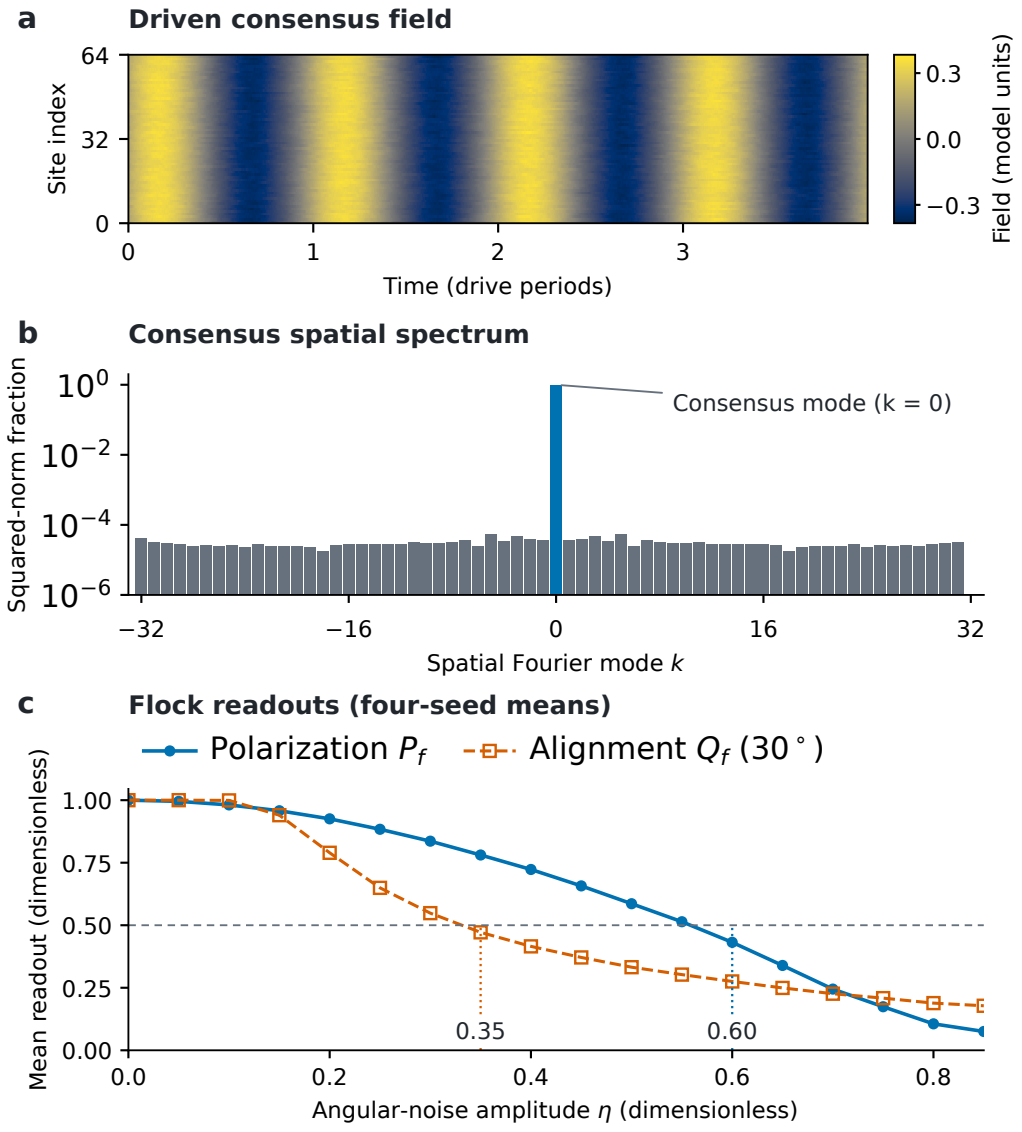


Figure 6: **Complementary physical readout and response measurements. a**, Delayed consensus on a 64-site degree-four ring over four drive periods, after subtracting each site's time average. Color shows signed field values in model units, with zero marked. A coherent oscillation dominates the transverse residual. **b**, Spatial Fourier fractions of the time-centered field's squared norm; the $k = 0$ mode dominates (log scale). These weights differ from the peak-amplitude and quadrature weights in Appendix E.4.1. **c**, External-field Vicsek flock: polarization P_f (blue circles) and alignment within $30°$, Q_f (orange hollow squares), versus dimensionless angular noise η. Hollow markers reveal coincident points without offsets. The dashed line marks one half; dotted guides mark the first sampled noise levels where alignment and polarization fall below one half, at $\eta = 0.35$ and $\eta = 0.60$, respectively. The readouts are numerically ordered on $[0.15, 0.70]$; the full sweep extends beyond it. The readouts are not nested measures, and the sweep is not a boundary-crossing trajectory. Panels a,b use one fixed seed at $\tau = 0.46$; Table 1 uses the separate 13-delay record. Panel c averages four seeds; the data record supplies no uncertainty intervals.

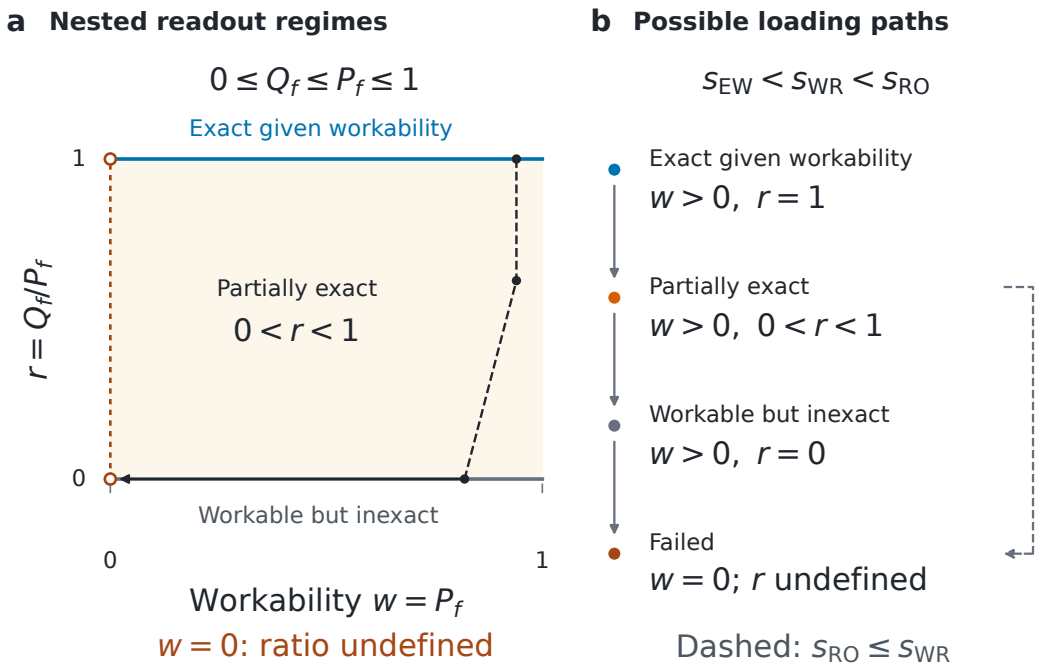


Figure 7: **Readout-regime map and boundary order under monotone loading. a**, The four local readout regimes in canonical reduced coordinates (w, r): one open partially exact region, the line $r = 1$ (exact given workability), the line $r = 0$ (workable but inexact), and the failed case $w = 0$. Hollow endpoints exclude $w = 0$ from the ratio chart: failure is specified without a value of r. The plotted path is schematic; measured coordinates are reported in Appendix E.3.2 and are not encoded by the path geometry. **b**, Along a specified monotone loading path whose boundary encounters cross between adjacent readout regimes, the four-step sequence holds when all three encounters occur with $s_{EW} < s_{WR} < s_{RO}$. An unencountered transition has first-hit parameter $+\infty$. If $s_{RO} < \infty$ and $s_{RO} \le s_{WR}$, the path can pass directly from the partially exact region to the failed boundary. Nested readouts fix the allowed boundary cases but do not determine their encounter order.