

Same physical state, different collective dynamics: state encodings select synchronization outcomes in language-model agents

Takahiro Ezaki^{1,*}, Naoto Imura¹, Katsuhiro Nishinari^{1,2}

¹Research Center for Advanced Science and Technology,
The University of Tokyo, Tokyo, Japan

²Department of Aeronautics and Astronautics, School of Engineering,
The University of Tokyo, Tokyo, Japan

*Correspondence: tkezaki@g.ecc.u-tokyo.ac.jp

Abstract

Language-model agents act on state encodings of their environment, yet these are treated as interchangeable interfaces. Using pretrained language models, we designed a circular-synchronization experiment applying a state-encoding intervention while holding the physical system fixed: each agent sees only a summary of its neighbours' relative phases and chooses to advance, stay or retard. Encoding that state as low-order circular moments rather than as a histogram selected different collective outcomes. In GPT the moment encoding synchronized the population in 6/6 seeds and the histogram encodings in 0/6; the effect replicated in Claude but reversed direction. Replaying identical fields shifted each agent's advance/stay/retard probabilities far beyond within-encoding repeat variation, in GPT, Claude and Gemini; in GPT, presentation alone shifted the operator with the moment values fixed. State encodings therefore form part of a model-dependent effective interaction law, not a neutral interface.

Language-model agents never act directly on an environment; they act on an encoded description of it. The same physical state may be supplied as a statistical summary, a histogram or a structured text record, usually fixed as an implementation choice in the agent scaffold. This transformation from physical state to model input is the state encoding of our title; in what follows we call it the *observation map*, and reserve *serialization* for how a fixed set of state variables is arranged as text. Because the model acts on the resulting string, alternate encodings need not induce the same effective policy, and LLM outputs are sensitive to prompt formatting¹, to the order² and position³ of information and to the labelling of choices⁴. Single-turn sensitivity of this kind is established; what is not is whether it survives feedback, so that a population driven by one encoding ends in a qualitatively different state than the same population driven by another.

That distinction matters because language-model agents are increasingly assembled into communicating populations⁵⁻⁸. Such populations form social conventions and amplify collective biases⁹, show network-dependent behaviour¹⁰, produce non-trivial outcomes in social dilemmas¹¹ and steer markets toward concentration¹²; critical analyses call for explicit measurement of purported emergent behaviour¹³.

What makes a population different from a single agent is feedback: a small change in a stochastic action distribution may disappear, accumulate or redirect the states agents observe later. This connects to performative and sequential distribution shift, in which a deployed policy changes the data on which it is later evaluated^{14,15}, and to the fact that policies act on observations rather than latent states, so

the state representation is part of the effective policy^{16,17}. Two questions follow: does an encoding change the microscopic *response operator* on the same state (the three probabilities the model assigns to advancing, staying and retarding on a given field), and does that difference survive feedback to change a macroscopic collective outcome?

We address these questions with a deliberately minimal circular synchronization assay. Phase-based oscillator models suit this purpose because they connect a microscopic interaction rule to interpretable macroscopic observables such as locking, partial order and collective-frequency shifts^{18–20}, across systems ranging from chemical oscillators²¹ and power networks²² to circadian clocks²³. We use synchronization as a controlled assay rather than a literal Kuramoto model: the interaction law is not prescribed as a sinusoid but measured as the stochastic action rule of a pretrained language model on an encoded relative-phase field. Each agent observes only its peers’ relative phases and chooses to advance, stay or retard; the deterministic engine alone applies the coupling, which the model never sees. We encoded each relative-phase field either by its first three circular moments or by a 24-bin histogram, serialized by bin centers or by bin intervals with the same mass in every bin: moments compress the field, whereas the two histogram encodings carry identical masses under different labels. To separate presentation from information content, we also rendered the same moment values in alternative layouts and with added task-irrelevant context.

Changing the encoding alone was sufficient to shift synchronization outcomes systematically, in opposite directions in GPT and Claude; identical-field replay traced the dependence to the microscopic response operator, which in GPT shifted even when only layout or context changed.

Results

Observation maps select distinct collective outcomes in GPT agents

We constructed a synchronous circular-agent system, a population of phase oscillators in the tradition of coupled-oscillator models^{18,20}. Agent i has an unwrapped phase $x_i(t)$, a wrapped phase $\theta_i(t) = x_i(t) \bmod 2\pi$ and a fixed natural increment ω_i . At each step a language model saw only a text description of the other agents’ phases relative to agent i and returned $f_i(t) \in \{-1, 0, +1\}$, with no goal and no instruction to synchronize. The deterministic engine then applied

$$x_i(t+1) = x_i(t) + \omega_i + K f_i(t). \quad (1)$$

Here K sets how far a single chosen action shifts the phase in one step. The model was not given K , absolute phase, agent identity, time, history or a preferred state. The three encodings were *moments*, the first three circular moments of the field, which summarize its mean direction and concentration at three harmonic orders; *centers*, a 24-bin histogram labelled by bin center; and *intervals*, the same 24 bin masses labelled by bin interval to six decimals.

This minimal design closes off the usual alternative explanations before the comparison is made. Each call is stateless: with no memory, identity, time or goal, the response can depend only on the encoded field, so the encoding-specific action rule is a well-defined object, and no encoding can benefit from learning within a run or from being asked to synchronize. The coupling lives only in the shared deterministic engine, so a difference between conditions cannot come from the physics. All three payloads are computed from the same 24-bin measurement of the same field, so what an encoding retains or discards is a property of the encoding itself, not of a noisier sensor. Finally, the sampled action is the only channel from model to engine, so anything that separates the conditions must pass

through the action distribution. Initial conditions and sampling luck remain, and are handled by matching seeds across encodings and by the exact $K = 0$ control below.

We first tested whether the encoding changed the closed-loop dynamics of GPT agents, using $N = 17$ agents over $T = 100$ steps at couplings $K \in \{-0.15, 0, 0.08, 0.15\}$, six physical seeds for the initial phases, matched across encodings within each coupling value, and one fixed natural-increment vector reused across all seeds, encodings and model families, and `gpt-5.4-mini`. Physical dynamics, action set and integrator were identical across encodings, and the instructions were fixed except for the line describing the encoding (Fig. 1a). All 122,400 calls returned valid actions, and the independent unit for every comparison below is the physical seed, not the agent, time step or call.

Positive coupling produced a qualitative separation. We summarize synchronization by the polar order parameter $r_1(t)$, which runs from 0 when phases are scattered to 1 when perfectly aligned (Methods, Eq. (8)). At $K = 0.08$ the mean final value was 0.999 for moments, 0.710 for centers and 0.382 for intervals (Fig. 1b–d); at $K = 0.15$, 0.996, 0.727 and 0.507. Under the criterion $r_1(T) \geq 0.9$, fixed before the outcomes were examined, moments locked in all six seeds at both positive couplings, whereas centers and intervals locked in none (Supplementary Figs. S3,S4). All six paired seeds favoured moments over each histogram encoding ($p = 0.03125$, two-sided exact sign test; Fig. 1c,d). Because the two nearby coupling values test the same prespecified contrast, we treat them as consistency checks on one qualitative result rather than as independent findings. Centers mostly produced partial alignment, whereas intervals more often retained two-cluster structure. A perfectly synchronized one-cluster state makes both r_1 and r_2 large, whereas two opposite clusters make r_2 large but r_1 small, so we use the descriptive contrast $Q_2 = r_2 - r_1$, which discounts ordinary one-cluster alignment and is what distinguishes two-cluster and higher-harmonic states^{24,25}; for intervals it averaged 0.128 at $K = 0.08$ and 0.213 at $K = 0.15$, against -0.004 and -0.013 for moments (Fig. 1c,e).

The $K = 0$ condition is an exact negative control: the coupling term in Eq. (1) vanishes, so all three encodings produced identical r_1, r_2, r_3 trajectories from each shared seed even though their action distributions still differed (Fig. 1f and Supplementary Fig. S1). Mismatched initial conditions and an encoding-specific engine are thereby excluded. At $K = -0.15$ all three stayed low in polar order but remained active, their mean signed actions (*social torques*; Methods, Eq. (9)) differing in sign, at -0.335 for moments, -0.033 for centers and $+0.062$ for intervals: suppression of polar order rather than inactivity (Supplementary Fig. S5).

Controlled fields elicit encoding-dependent response operators

So far the encoding and the feedback are entangled: each encoding drives the system along a different trajectory and so presents a different sequence of fields. To measure the response operator alone, we presented synthetic fields whose shape we controlled directly, rendered under each encoding as shown in Fig. 2a, and estimated for each field ρ and encoding $\mathcal{R}$ the probability of the three actions from repeated queries,

$$p_{\mathcal{R}}(\rho) = [p_{\mathcal{R}}(-1 \mid \rho), p_{\mathcal{R}}(0 \mid \rho), p_{\mathcal{R}}(+1 \mid \rho)], \quad (2)$$

from which we derived the activity $A_{\text{op}}(\rho) = 1 - p_{\mathcal{R}}(0 \mid \rho)$ and the signed mean action $a_0 = p_{\mathcal{R}}(+1 \mid \rho) - p_{\mathcal{R}}(-1 \mid \rho)$. We also rotated one single-peaked field around the focal agent and recorded the mean action at each angle. Decomposing that curve by how often it repeats around the circle separates attraction or repulsion towards the peers, carried by the one-cycle component, from a two-lobed response compatible with opposed groups. We quantify these by Fourier coefficients (Methods), a data-driven counterpart of the phase-interaction functions of coupled-oscillator theory²⁶, reconstructed

experimentally from real oscillators²⁷.

We first varied how tightly the single peak was concentrated, through concentration values $\kappa \in \{2, 4, 6, 9, 12\}$, sampling the rotation angles at slightly irregular, mirror-paired positions so that a fast component could not masquerade as a slow one (25,920 valid calls; Methods). Sharpening the same field did not move the three encodings together: moments ended as a purely directional controller that always acted, intervals as a two-lobed response with little net direction, and centers as a reversed directional response with a net retarding bias (Supplementary Figs. S6–S8; Fig. 2b shows the three full action distributions on one such field). Each characterization was specified in advance and had to survive being measured twice: the experiment was collected as two separately acquired batches, the *acquisition blocks*, each fitted on its own, and all three reproduced (Supplementary Figs. S6–S8). The differences were not confined to one coefficient. The total-variation distance measures how far apart two advance/stay/retard distributions are, on a scale where 0 is identical and 1 no overlap. Across two-peaked, asymmetric, antipodal and sparse few-peer fields, the encodings continued to differ in activity, signed bias and that distance, even where the one-cycle component a_1 that produces attraction, and hence locking, was weak (Fig. 2c–g and Supplementary Figs. S9,S10). An antipodal field, in which peers split into two opposite groups, is symmetric about the focal agent, so a_1 collapsed for all three encodings; yet they still differed in the full action distribution, with moments switching from always acting to mostly abstaining (Fig. 2c). No tested stimulus class brought the three encodings together (Fig. 2g). Sweeping the directional imbalance ε , the signed excess of peers on one side, moments was almost inactive at balance but acted already at the smallest nonzero imbalance tested, $|\varepsilon| = 0.02$ (Fig. 2d–f): sharp, but grid-limited rather than discontinuous.

Replay on identical endogenous fields isolates an operator effect

The decisive control was an identical-field replay, which re-encodes fields the closed loop actually produced, so any response difference can come only from the encoding. We call the encoding of the trajectory that originally generated a field its *source* encoding, and the encoding used to show that frozen field during replay its *presented* encoding. Using a rule fixed in advance, we selected 48 fields that arose during the collective runs, balanced across six trajectory-state types and across source encoding (Supplementary Figs. S11–S13). Each was re-encoded and presented to GPT under all three presented encodings in turn, for 32 responses in two blocks (4,608 valid calls; Fig. 3a).

The same field did elicit different action distributions, separated by a mean pairwise total-variation distance of 0.344. Permuting the presented-encoding labels within each field, none of 5,000 relabellings produced a separation this large (Fig. 3f; $p = 0.0002$, the resolution limit; Methods). Resampling the 48 fields as whole units, all three pairwise intervals lay entirely above the test–retest floor, so no single pair carried the result. The separation was 3.76 times the test–retest variation between the two acquisition blocks of the same encoding on the same field (0.344 versus 0.092; Fig. 3e, and *block noise* where the figures name it). Fig. 3d gives the same comparison pair by pair: centers and intervals are information-matched, differing only in how the same 24 bin masses are labelled, and they separated by 0.290 (95% CI 0.220–0.364), 3.2 times that block noise, so serialization moves the operator with the retained information fixed.

At a fixed field the encoding is the only changed input and the sampled action the only output entering the engine, so any causal path from encoding to dynamics must pass through the action distribution. Two further prespecified tests asked whether that operator effect is itself field-dependent: whether fields generated under different encodings draw different average actions (Fig. 3c; $p = 0.078$),

and whether re-encoding’s effect depends on which encoding generated the field (Fig. 3b,g; $p = 0.072$). Neither test crossed the prespecified threshold, so source-dependent modulation remains unresolved; component estimates are reported in Supplementary Fig. S13g.

The collective encoding effect replicates in Claude, but reverses direction

Model families and multi-agent systems can behave differently across comparable evaluations^{28,29}, so we asked whether this dependence was family-specific. On the same panel we collected 16 responses per presented encoding from `claude-haiku-4-5-20251001` and `gemini-3.5-flash` (2,304/2,304 valid for Claude, 2,281/2,304 for Gemini, invalidity not concentrated in any encoding). Re-encoding the same field again changed the action distribution in both ($p = 0.0002$ for each). Relative to test–retest variation the effect was larger in Claude and Gemini than in GPT, but mainly because both answer near-deterministically and so repeat themselves more closely, not because the encoding moved them further; the separations, floors and ratios for all three families are given with Fig. 4f and Supplementary Fig. S14.

Microscopic replication does not imply that an encoding selects the same collective outcome in every model, so we ran a second matched collective experiment in Claude, changing only the model family (Fig. 4a) and reusing the six GPT seeds at $N = 17$, $T = 100$ and $K \in \{0, 0.08, 0.15\}$; at $K = 0$ the trajectories again coincided exactly. At positive coupling the trajectories separate by encoding as in GPT, but not in the same order (Fig. 4b).

The two positive-coupling outcomes were combined within a prespecified seed-index block, so that the two coupling values were not treated as independent replications: a binary score for terminal locking, whether a run reaches polar lock and holds it to the end, and a continuous score averaging final polar order (Methods, Eqs. (12) and (13)). Comparing the observed separation with every reassignment of the three encoding labels within each seed, keeping the two couplings together, it was unusual on both scores ($p = 0.00103$ and $p = 0.000386$ over $(3!)^6 = 46,656$ reassignments), with means ordered moments < centers < intervals (Fig. 4d). At $K = 0.15$ moments locked in 0/6 seeds and the histogram encodings in 6/6; at $K = 0.08$ the counts were 0/6, 2/6 and 5/6 (Fig. 4c). Four further seeds, acquired after the confirmatory design was fixed, enter none of the reported p -values and only checked that the direction reproduced. Pairwise contrasts supported moments versus each histogram encoding but did not order intervals against centers. Both families therefore showed a collective encoding effect, but the encoding-to-locking relationship reversed (Fig. 4e and Supplementary Figs. S15–S17): moments produced synchronization in GPT but only partial order in Claude, and the histogram encodings the opposite. None of the three tested encodings was uniformly optimal across GPT and Claude.

Presentation changes alter the operator when task-relevant information is fixed

The three encodings differed in how much of the field they retained and in how it was laid out. To ask whether retained information alone drove the effect, we built three GPT inputs carrying the same circular-moment values presented differently: the original text, the same numbers as a table, and a version with added task-irrelevant text, compared on the same 48 replay fields with 16 responses per version in two blocks (2,304 valid calls; Fig. 5a and Supplementary Fig. S18).

Even with identical numbers, presentation mattered: the mean pairwise distance among the three versions was 0.311 ($p = 0.0002$; Fig. 5b–d), against a within-version block noise of 0.069. Re-laying out the numbers as a table moved the operator little, by 0.145 or 2.1 times that noise, and its interval (0.079–0.219) overlapped the test–retest range (0.049–0.090). Adding task-irrelevant context moved it

by 0.414, exceeding the 0.344 between different encodings (Fig. 5c,e,f).

The padded version changed context volume, the position of relevant numbers³ and length together, so it tests none of them cleanly^{1,2,30,31}. In two secondary controls, we examined a simple prompt-length explanation. Holding one moment's prompt fixed, a monotone length account predicts that whichever histogram prompt lies farther in characters is also farther in response; the opposite happened at both anchors, and a second control found almost no change in response distance when the character gap was multiplied by 2.6 to 7.8 (Supplementary Figs. S19,S20).

Discussion

Our results identify the observation map as a constituent of the effective policy a language-model agent implements. The same relative-phase state elicited different action distributions under moments, bin centers or bin intervals, these differences survived on fields the agents generate themselves¹⁴, and changing the map alone altered collective order in two model families. This is stronger than an open-loop prompt-sensitivity result, because the interface change propagated through feedback to select different outcomes. The cross-family reversal is central: the simple reading of the GPT experiment, that moment compression is intrinsically synchronization-promoting, is falsified by Claude under the same design, so the outcome depends on the model and observation-map pair, not on a superior encoding. Changing either need not preserve the effective interaction law, so validation should not transfer by analogy, consistent with broader critiques of multi-agent evaluation^{13,29}. This mirrors a classical point in sequential decision-making, that the observation model is part of the policy^{16,17,32,33}, and extends to systems where agent populations shape market-level outcomes^{12,34–36}.

The identical-field replay locates that dependence microscopically, in GPT, Claude and Gemini alike. Since the sampled action is the only model output entering the engine, the action channel is exhaustive by design; what remains open is whether the fixed-field differences measured here are quantitatively sufficient to reproduce the encoding-specific trajectories, which an operator swap at a fixed state sequence would test.

In GPT, the encoding effect was not explained by the task-relevant numerical values alone: reformatting the same moments changed the response distribution^{1,2}, and adding task-irrelevant context produced an effect comparable to the main contrast. In a secondary, post hoc crossed analysis of four histogram serializations, response differences tracked explicit coordinate binding (whether each bin's mass is written together with its coordinate) more closely than character-count differences^{30,31}; this does not establish a general serialization mechanism.

We also asked whether a cheap statistical model trained on controlled fields could stand in for the language model inside a collective simulation, as learned models replace costly dynamics elsewhere^{37–39}. Ordinary cross-validation tests new fields from the same acquisition; deployment asks whether the model still holds on the fields an interacting population generates. All three met their branch-specific in-domain criteria, but only the revised moments model cleared the closed-loop support stage, the check that the fields the interacting population actually generates are covered by mutually consistent training data, and was evaluated prospectively. For centers and intervals the nearby training fields either failed to reproduce the collective-field response or disagreed with one another, so both branches stopped under their prespecified rules before any prospective acquisition (Supplementary Figs. S21–S25). It is secondary, but reinforces the same conclusion: agent approximations must be validated on the state distribution the closed loop generates^{14,15,40–42}, a stronger requirement than robustness to exogenous shift⁴³.

Several limitations bound the conclusions. The collective experiments used a minimal, stateless, three-action system at a single size $N = 17$, with a small coupling grid and a fixed horizon, so the phenotypes are finite-system outcomes, not a phase transition or a universal critical coupling^{20,44}. Collective dependence was tested in GPT and Claude, the presentation controls only in GPT, and all three tested models were lower-cost, non-frontier offerings queried under one prompt contract at a fixed temperature, so neither frontier-scale models nor the temperature dependence of the operator is addressed. Backend updates⁴⁵, memory, richer actions and task environments may alter the map.

They also suggest the next tests. A further collective experiment would test whether the reversal recurs in another model family, and operator-swap experiments could close the quantitative gap above. The principle should then be tested in task-based multi-agent settings, where coordination architectures and message-passing interfaces are explicit design choices^{5-7,46,47}. Practically, the observation serializer is a versioned component: it belongs in what an agent evaluation reports, and must be revalidated with the model, in the closed loop in which both will operate.

Methods

Statistical analysis overview

The experiments below ask different questions and therefore have different units of inference. The table states, for each analysis, what is asked, which unit is permuted or resampled, what the repeated model calls contribute, and what is compared.

Table 1: Units of inference by analysis. For each experiment, the question asked, the unit that is permuted or resampled, what the repeated model calls contribute and the comparison that carries the result.

Analysis	Question	Unit of inference	Role of repeated calls	Main comparison
GPT collective	Does the encoding change the trajectory?	physical seed ($n = 6$)	generate the trajectory	paired within seed
Controlled fields	Does the response rule on a fixed field change?	field and acquisition block	estimate action probabilities	response curve and its harmonics
Identical-field replay	Does the operator change on the same physical field?	physical field ($n = 48$)	estimate the trinomial per field	encoding labels swapped within field
Claude collective	Is there a macroscopic effect in Claude?	confirmatory seed-index block ($n = 6$)	generate the trajectory	all label permutations within seed
Same-information control	Does presentation matter at fixed numbers?	physical field ($n = 48$)	estimate probabilities per variant	variant labels swapped within field
Surrogate	Can responses be predicted in and out of distribution?	stimulus profile or complete run	provide training labels	out-of-fold and prospective

Model calls are not replicates: repeated calls estimate the action distribution at a fixed unit, and permutations and bootstraps keep together all repeated measurements belonging to the same unit. Unless stated otherwise, reported confidence intervals are percentile cluster bootstraps with 5,000 resamples, drawing all of a unit’s values together.

Unless stated otherwise, a confidence interval reported here describes the variation across the physical seeds, physical fields or runs that were sampled in the corresponding experiment. Except for factors that were varied deliberately, such as the model family in the cross-family replications and the presentation form in the controls, these intervals carry no uncertainty from future provider updates, from other prompt wordings or from other model families. Confirmatory inference in this study rests on exact sign tests and exhaustive within-seed permutations and does not depend on any bootstrap or normal-approximation interval.

Protocols, observation maps, prompt text, model IDs, physical seeds, analysis endpoints and stopping rules were versioned and hash-locked before the corresponding acquisitions, and *prespecified* is used throughout in that internal sense: the ordering is documented by our own versioned records rather than attested by a public registry. Raw responses, parser status and retry history were retained, and no failed response was silently converted into a valid action. What was locked and when, and the structural safeguard on replay-field selection, are given in the Supplementary Information; the hash manifest and the acquisition-level operational records are released with the code repository.

Circular-agent dynamics

Agent $i \in \{1, \dots, N\}$ had an unwrapped phase $x_i(t) \in \mathbb{R}$ (accumulating without the modulo- 2π reduction) and wrapped phase $\theta_i(t) = x_i(t) \bmod 2\pi$. All agents updated synchronously according to Eq. (1), a discrete-time circular dynamics in the tradition of coupled phase-oscillator models^{18–20}. Initial phases were generated from fixed, encoding-independent physical seeds. Natural increments were instead given by the deterministic, zero-mean vector of N evenly spaced values on $[-0.05, +0.05]$ rad per step, assigned to agents in fixed index order and reused unchanged across all seeds, encodings and model families; they therefore contribute no seed-dependent variation. The spread is small relative to the couplings by design: locking is feasible at both positive couplings because only the time-averaged action, not each discrete action, enters the phase balance. Without coupling the same spread separates phases by $(\max_i \omega_i - \min_i \omega_i)T = 10$ rad over the fixed horizon, so the $K = 0$ arm is far from synchronization by construction. Initial phases were drawn uniformly from $[-\pi, \pi)$ using a *physical seed*, a deterministically derived integer kept separate from the local seeds that controlled sampling order. Initial phases were thereby matched across encodings and model families within each coupling value, but were generated separately for different coupling values through a K -dependent seed. The action mapping was **retard** $\mapsto -1$, **stay** $\mapsto 0$ and **advance** $\mapsto +1$.

Negative coupling was implemented only through the sign of K in the engine. The model received the same instruction at positive, zero and negative K . At $K = 0$, the engine satisfies $x_i(t) = x_i(0) + t\omega_i$ independently of the sampled actions.

Relative-phase field

For focal agent i , peer j contributed

$$\delta_{ij}(t) = \text{wrap}[\theta_j(t) - \theta_i(t)] \in [-\pi, \pi). \quad (3)$$

The focal agent was excluded. Histogram masses were normalized by $N - 1$. The raw peer count, absolute phase, agent identity, time step, natural frequency, coupling and trajectory history were not included in the prompt. The encoder used a deterministic half-open boundary convention and a fixed field ordering, and it aggregated peers in a way that does not depend on the order in which they are

listed, so that the same multiset of relative phases always produces the same encoding. We verified two invariances of the encoder numerically: that rotating all phases by a common angle shifts the encoding consistently, and that permuting the peers leaves it unchanged. These were deterministic checks on the encoder code and involved no additional language-model queries.

Observation maps

The first three circular moments compactly summarize one-lobed, two-lobed and three-lobed angular structure (a net direction, a pair of opposed groups, a three-group arrangement) while discarding finer bin-level detail. The three principal maps were:

- (i) **Moments:** the real and imaginary components of the binned circular moment

$$\tilde{z}_m = \sum_b h_b \exp(imc_b), \quad m = 1, 2, 3, \quad (4)$$

serialized in a fixed narrative format. The production feature is the binned moment $\tilde{z}_m$, i.e. the m -th moment of the fixed 24-bin representation (bin masses h_b , bin centers c_b); for finite-peer fields it approximates the raw peer moment $z_m = \frac{1}{N-1} \sum_{j \neq i} \exp(im\delta_{ij})$. All three maps are therefore deterministic functions of the same binned field, and the moments agent received $\{\tilde{z}_m\}$ rather than the raw peer moments.

- (ii) **Centers:** 24 normalized bin masses h_b , each paired with its bin center c_b , in a fixed canonical order.
- (iii) **Intervals:** the same 24 masses paired with half-open bin intervals and serialized to six decimal places.

Complete prompts, token counts under each provider tokenizer and boundary conventions are reported in Supplementary Fig. S2 and Supplementary Table S1.

Language-model response contract

Each backend call was stateless and consisted of a single user message requiring the one-field JSON object `{"social_action": "advance|stay|retard"}`. The fixed instruction (prompt contract `response-law-v0.1`) was identical across maps except for the one-line observation description and the payload heading that labels it, and is reproduced in full in Supplementary Fig. S2. Responses were read by a fixed two-stage deterministic parser: strict parsing of that object, and, on failure, a rescue rule that reads only an action word the model itself wrote and never assigns, changes or defaults an action. A response from which no action word could be read was recorded as invalid and retried, up to three attempts. No invalid response entered the retained collective runs, and unrecoverable responses were excluded rather than defaulted; probabilities were normalized over valid calls. The parser stages are strongly family dependent; per-family and per-encoding dispositions are given in Supplementary Table S2 and released with the code repository.

The three model IDs were `gpt-5.4-mini`, `claude-haiku-4-5-20251001` and `gemini-3.5-flash`. Generation parameters, provider and API endpoint, acquisition date range, SDK/API versions, sampling parameters (including parameters left unset, for which the provider default applied), retry policy and provider seed support are reported in Supplementary Table S2.

Microscopic response acquisition

For a fixed physical field ρ , repeated model calls estimated the operator in Eq. (2). The point of this sweep is to obtain, for a language-model agent, the object that coupled-oscillator theory calls a phase interaction function: the dependence of the interaction on relative phase alone, which weakly coupled oscillators reduce to²⁶ and which has been measured in real oscillator experiments²⁷. In that theory the odd part of the function governs attraction and hence locking, the even part shifts the collective frequency, and harmonics beyond the first admit multi-cluster states, so resolving $g_{\mathcal{R}}$ by harmonic places a text-level manipulation on the same axis as an interaction law. The sine coefficient a_1 is the odd first-harmonic component and therefore the analogue of the Kuramoto coupling strength. Because $g_{\mathcal{R}}$ is a signed mean, it is blind to changes that leave the mean action unchanged while redistributing probability between acting and abstaining, which is why the full trinomial and A_{op} are reported alongside it. For a single-peaked field rotated around the focal agent by an angular offset δ , so that ρ_{δ} denotes the field translated by δ , the signed response curve, a data-driven analogue of a phase coupling function^{26,27}, was

$$g_{\mathcal{R}}(\delta) = p_{\mathcal{R}}(+1 \mid \rho_{\delta}) - p_{\mathcal{R}}(-1 \mid \rho_{\delta}). \quad (5)$$

We fitted

$$g_{\mathcal{R}}(\delta) = a_0 + \sum_{m=1}^M [a_m \sin(m\delta) + b_m \cos(m\delta)], \quad (6)$$

where a_m multiplies the sine and b_m the cosine, so that a_m , not b_m , is the odd component here. The complex coefficient was $C_m = a_m + ib_m$, with $R_m = |C_m|$ and $\phi_m = \arg C_m$. The quantity fitted was the mean signed action at each of the 36 offsets, not the individual calls, and the fit was carried out separately within each acquisition block. Uncertainty on (a_m, b_m) came from a multinomial bootstrap within each offset, and a phase was reported only where its amplitude was large enough and its uncertainty region excluded the origin.

The primary concentration sweep used the three encodings, five concentration values, 36 translation offsets, 24 responses per cell and two separate acquisition blocks. The 36 offsets were mirror-paired and slightly jittered rather than evenly spaced, preventing higher harmonics from aliasing onto the low-order components used for interpretation. A Fourier order of $M = 6$ was the primary fit, with $M \in \{2, 4, 8, 12\}$ as prespecified sensitivity analyses. The analytic validation of the offset design and the recovery tests are reported with Supplementary Fig. S6 and in the Supplementary Methods.

Implementation conventions, uncertainty construction and sensitivity analyses are reported in the Supplementary Methods.

Controlled stimulus extensions included bimodal, asymmetric, exact antipodal, signed-imbalance and sparse finite-peer fields. Here the signed imbalance ε is the net directional excess of peers on one side of the focal agent (with $\varepsilon = 0$ a perfectly balanced field); the dense grid was

$$\varepsilon \in \{-0.10, -0.05, -0.02, 0, 0.02, 0.05, 0.10, 0.20\}. \quad (7)$$

Activation scales inferred from this grid were treated as resolution-limited.

Collective observables and operational phenotypes

Order parameters, the standard Kuramoto measures of collective coherence^{20,26}, were

$$r_m(t) = \left| \frac{1}{N} \sum_{i=1}^N \exp[i m \theta_i(t)] \right|, \quad m = 1, 2, 3, \quad (8)$$

so that r_1 measures alignment on the circle (phase locking) and r_2 measures two-cluster structure. We further used $Q_2 = r_2 - r_1$, the realised collective activity $A_{\text{run}}(t) = N^{-1} \sum_i \mathbf{1}[f_i(t) \neq 0]$ (distinct from the expected operator activity A_{op} above), and the full-run social torque

$$\tau_{\text{run}} = \frac{1}{NT} \sum_{t=0}^{T-1} \sum_i f_i(t), \quad (9)$$

the run-averaged net action per agent (positive when agents advance and negative when they retard). With phase states indexed $t = 0, \dots, T$ and actions sampled at $t = 0, \dots, T-1$, summing Eq. (1) over t yields the exact engine identity

$$\Omega_{\text{coll}} = \frac{1}{N} \sum_i \frac{x_i(T) - x_i(0)}{T} = \bar{\omega} + K \tau_{\text{run}}, \quad \bar{\omega} = \frac{1}{N} \sum_i \omega_i, \quad (10)$$

The population's mean drift therefore departs from $\bar{\omega}$ by exactly K times the social torque. Windowed torques (for example the final-20-step average τ_{20}) are named explicitly; reported torques are run-averaged unless stated otherwise.

The prespecified endpoint labels were:

- polar locked: $r_1(T) \geq 0.9$;
- high- r_2 non-polar: $r_2(T) \geq 0.5$ and $r_2(T) > r_1(T)$;
- partial polar order: $0.35 \leq r_1(T) < 0.9$, excluding high- r_2 ;
- low-polar active: $r_1(T) < 0.35$.

These labels provide operational finite- N phenotypes with qualitative reference points in higher-harmonic oscillator dynamics^{24–26}; they are not stability classifications or thermodynamic phases. The endpoint label uses the non-strict criterion $r_1(T) \geq 0.9$, whereas *terminal lock* uses the strict criterion $r_1(s) > 0.9$: a run showed terminal lock if there existed a time t^* with $r_1(s) > 0.9$ for every $s \in [t^*, T]$, and the terminal-lock time was the smallest such t^* . It admits $t^* = T$ (no minimum duration); behaviour under an imposed minimum duration is reported in Supplementary Fig. S4. Both thresholds were fixed before confirmatory analysis.

GPT matched collective experiment

This matched collective (macro) experiment in GPT is the one summarized in Fig. 1; it used $N = 17$, $T = 100$, four coupling values and six seed indices per coupling, with initial phases matched across encodings within each coupling value. Each of the three encodings contributed 40,800 valid calls, for 122,400 in total. The local seeds that control call ordering and retries were encoding-specific and deterministically derived. Run order and task submission were fixed in advance to avoid any association between encoding and wall-clock order.

The primary locking comparison used the paired physical seed as the unit of analysis. Six seeds were fixed by acquisition cost, and an exact two-sided sign test on six paired differences has a floor of $p = 0.03125$, so this comparison is powered to detect unanimous effects only. For each positive

K , all six differences between moments and each histogram encoding had the same direction; the exact two-sided sign-test value was $2/2^6 = 0.03125$. Continuous paired effects were summarized with seed-block percentile bootstrap intervals: within each (coupling, contrast) cell the seed-level differences were resampled 2,000 times and the percentile interval was written to the seed-block summary table. The error bars drawn in Fig. 1d–f are percentile bootstrap intervals of the same kind, resampling the six physical seeds 5,000 times, drawn beside the individual seed values rather than in place of them. Agents and time points were not treated as independent replicates.

Identical-field replay

The replay-field selection algorithm was fixed before the presented-encoding outcomes were acquired. Candidate frames were tagged into six prespecified strata, deduplicated by a physical 24-bin histogram hash and selected by a source-by-stratum quota. The final panel contained 48 fields, 16 from each source map and eight per stratum, with at least one source contribution to every stratum.

Each field was encoded under all three presented maps and queried 32 times in two blocks. The primary presented-encoding statistic was the mean, over physical fields, of the three pairwise total-variation distances between action operators,

$$d_{\text{TV}}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \sum_{f \in \{-1, 0, +1\}} |p_f - q_f|. \quad (11)$$

Presented-encoding labels were permuted at the level of individual responses within a physical field, and the three operators re-estimated from the relabelled responses; the field is therefore the exchangeability block, not the unit being relabelled: labels are shuffled only within a field, never across fields. As a parallel check, a multinomial deviance statistic (twice the log-likelihood ratio between a presented-encoding-specific and a presented-encoding-pooled multinomial fit to the three action counts per physical field, summed over fields) was evaluated under the same field-blocked presented-encoding-label permutation and yielded the same decision as the total-variation statistic. The second prespecified global test, the source main effect, used the prespecified contrast in presented-encoding-averaged signed action per physical field, with source labels permuted within the prespecified stratum $\times$ coupling blocks; the third, the source-by-presented-encoding interaction, used the analogous field-level activity contrast. Each of the three is an omnibus test: it asks whether the levels of one factor differ at all, in the manner of a one-way analysis of variance, rather than which pair differs. All three were Monte Carlo permutation tests with 5,000 resamples, prespecified at $\alpha = 0.05$ and reported without correction across the three; because each carries a single α , the pairwise decompositions are effect sizes rather than additional tests. Permutation p -values reach a resolution floor of $1/5001 \approx 0.0002$ (plus-one estimator: the observed statistic is counted among the resamples, so no reported p is zero). Pairwise confidence intervals are field-cluster bootstraps.

Every field–presented-encoding cell was acquired twice, in two independent blocks, so the distance between the two blocks of the same cell measures how far the operator estimate moves on repeat measurement alone. This within-presented-encoding noise floor, referred to as *block noise* in the figures, was the between-block TV for the same presented encoding and physical field; it plays the role of a test–retest floor against which between-presented-encoding distances are read, and its comparison with between-presented-encoding TV was paired by field. The intervals reported for the noise floor and for the between-presented-encoding mean are likewise field-cluster bootstraps.

Cross-family microscopic replication

Claude and Gemini were evaluated on exactly the same 48 physical fields and presented encoders. Each family used 16 responses per field and presented encoding in two blocks, for 2,304 attempted calls per family. The prespecified replication gates were: a global presented-encoding permutation $p < 0.05$; between-presented-encoding TV greater than within-presented-encoding TV; a positive field-bootstrap interval for the difference; at least two of three encoding pairs above the noise floor; and an acceptable, non-skewed invalid-response rate. Exact pairwise ranking across model families was not required.

The $n = 16$ design was a prespecified conservative response count, not a per-family power optimum.

Claude matched macro replication

The Claude confirmatory core reused the six GPT physical seeds at $N = 17$, $T = 100$ and $K \in \{0, 0.08, 0.15\}$. Four new physical seeds were acquired only at $K \in \{0.08, 0.15\}$ as a held-out directional extension. The held-out seeds were not pooled into the confirmatory exact tests.

For each core seed s and encoding $\mathcal{R}$, the two positive couplings were combined into a terminal-lock score

$$L_{s\mathcal{R}} = \frac{1}{2} \sum_{K \in \{0.08, 0.15\}} \mathbf{1}\{\text{terminal polar lock}\} \quad (12)$$

and a continuous final-order score

$$Y_{s\mathcal{R}} = \frac{r_1(T; 0.08) + r_1(T; 0.15)}{2}. \quad (13)$$

For each core seed, the three encoding labels were permuted jointly across the two positive couplings. We enumerated all 46,656 within-seed permutations and computed the sum of squared deviations of encoding means for L and Y . Pairwise contrasts used seed-level bootstrap intervals and exact sign-flip tests (flipping the sign of each seed’s paired difference in all 2^6 ways); the bootstrap resampled the paired differences of the six core seeds 5,000 times and reported the percentile interval. The error bars drawn in Fig. 4d are seed-bootstrap percentile intervals of the same construction; the four held-out seeds are shown but enter neither the intervals nor the reported p -values.

What would count as a successful replication in Claude was fixed before the Claude acquisition as three ordered criteria, written as go/no-go gates so that a partial replication could not be reported as a full one. Gate A asked the weakest question, whether the encoding has any effect at all, and required global encoding effects for both terminal lock and continuous final order; Gate B asked whether that effect is qualitative and required a lock/nonlock separation rather than a graded shift; Gate C asked the strongest question, whether Claude reproduces the GPT map from encoding to phenotype. Gates A and B passed. Gate C failed; the observed result instead matched the outcome prespecified as the reverse map, which is a distinct prespecified alternative and not a failure of Gates A or B.

Same-task-information control

The control used the 48 replay fields and three variants of the moments observation: original, reformatted, and a task-irrelevant-padding variant with added task-irrelevant context. All task-relevant numerical moment values were identical across variants. Each field-variant cell had 16 responses in two blocks. The primary statistic and field-blocked permutation followed the replay presented-encoding analysis. Pairwise means, confidence intervals and the quoted noise-floor intervals are field-cluster bootstraps.

Character and token lengths were constant within each encoding and variant under the fixed-

width formatting and backend tokenizer. The task-irrelevant-padding condition was interpreted as a deterministic compound presentation manipulation, one that jointly changed context volume, the position of relevant values and length, rather than an isolated prompt-length intervention.

Serialization controls

Two supplementary controls held the retained numerical information fixed while varying prompt length, layout and whether each mass was explicitly bound to a bin index or coordinate, on the same 48 replay fields (3,072 valid calls of 3,072 in each control). Their prespecified contrasts were not diagnostic; interpretation therefore relies on secondary, post hoc comparisons within the crossed designs. Full constructions and results are reported with Supplementary Figs. S19 and S20.

Surrogate compressibility and transportability

This analysis is secondary to the observation-map intervention and its results are reported in the Supplementary Information only. It asks a practical question: a population of $N = 17$ over $T = 100$ steps costs 1,700 model calls, and a related strategy in model-based learning is to replace costly dynamics with a learned surrogate^{37–39}. The object being fitted is the same three action probabilities the replay measures: given a field, predict the probability of **retard**, **stay** and **advance**. The inputs are descriptors of the physical field, not the text the model saw, and they exclude coupling, time and source-encoding labels for the same reason the agent is never given them: a surrogate fitted without them can later stand in for the agent under exactly the information the agent has. The centers and intervals surrogates receive the same descriptors, so any difference between those two branches comes from the responses the language model produced under the two serializations and not from the feature space.

Formally, a field ρ enters the surrogate only through its prespecified descriptor vector $x(\rho)$. These descriptors describe the physical field itself and are computed from the same 24-bin record in every branch; they are not the prompt any encoding presented. The branch name (moments, centers or intervals) says only which encoding’s measured responses supply the training labels: the moments branch uses the 15 descriptors common to all branches, and the two histogram branches additionally receive the 24 bin masses (39 inputs in total), so that the bin-level detail their prompts carried is also available to their surrogates. What is *measured* is the language model’s empirical action distribution $\hat{\mathbf{p}}(\rho) = (\hat{p}_-, \hat{p}_0, \hat{p}_+)$, estimated from $n(\rho)$ repeated calls with action counts $c_f(\rho)$, $f \in \{-1, 0, +1\}$. What is *predicted* is a distribution $\mathbf{q}_\theta(x(\rho))$ over the same three actions, fitted by minimizing, over the set of measured fields $\rho_1, \dots, \rho_M$ entering the fit, the multinomial log loss

$$\ell = - \frac{\sum_{i=1}^M \sum_{f \in \{-1, 0, +1\}} c_f(\rho_i) \log q_{\theta, f}(x(\rho_i))}{\sum_{i=1}^M n(\rho_i)}, \quad (14)$$

where θ collects the parameters of whichever candidate model is being fitted (the candidate classes are listed in the Supplementary Methods), and ℓ is the mean negative log-likelihood per model call (in nats; lower is better); the same ℓ evaluated on held-out data is the reported score. What every judgement is *based on* is a comparison of losses: the reference is the peer-count baseline $\bar{\mathbf{q}}^{(m)}$, the average training response at peer count m , which ignores the shape of the field entirely, so a surrogate is credited only

where $\Delta\ell = \ell_{\text{baseline}} - \ell_{\text{model}} > 0$, that is, only where the shape of the field carries usable information.

We kept three questions separate, because passing one does not imply the next: in a closed loop the model’s own actions change the fields it will be asked about, so predicting well on the fields used for training does not show that the surrogate can be used inside a simulation. *Compressibility* asks whether the surrogate predicts the response to controlled fields it has never seen: the data are split into folds, each fold is hidden in turn, the model is fitted on the rest, and Eq. (14) is evaluated on the hidden fold only (cross-validation), always keeping every measurement of one physical field inside the same fold and judging the result against the baseline. *Closed-loop support* starts from the fields that actually arose in the language-model collective runs. For each such field, the training set is searched for the physically most similar fields: the nearest-neighbour distance d_{NN} scores whether any sufficiently similar training field exists at all, and the local response dispersion V_{local} scores whether the language model’s measured responses on those similar fields agree with one another; both are held to prespecified thresholds, and a field failing either would force the surrogate to guess beyond its evidence. *Transportability* puts the frozen surrogate in the language model’s place: starting from the same initial condition, with the same engine, coupling and horizon, the surrogate generates its own complete closed-loop trajectory, and its run-level endpoints E (action rates and collective order) are compared with those of a new language-model run through the per-run absolute error $|E_{\text{pred}} - E_{\text{obs}}|$; it is reported only where such new runs were acquired. The fold construction, the model-selection split, the definitions and thresholds of d_{NN} and V_{local} , and the endpoint list are given in the Supplementary Methods.

All three encodings met their prespecified, branch-specific in-domain criteria. Centers and intervals stopped at the closed-loop support stage under their prespecified rules, and so were never evaluated prospectively rather than measured and found wanting. For moments, checking the fitted surrogate against the measured replay responses on the fixed fields of Fig. 3 exposed a systematic error confined to one type of field: on active collective fields the surrogate kept predicting that agents would hold still when they in fact moved, whereas on fields whose neighbours are symmetric about the agent, where the measured operator really does abstain, its predictions were correct. A revision that separates those two cases using field descriptors alone was frozen and then evaluated prospectively against 44 new collective runs, where it improved average prediction of microscopic actions and of final collective order without being uniformly superior on every endpoint. These conclusions apply to the model families, descriptors and acquisition domains tested. The descriptor list, candidate model families, fold definitions, baseline, distance and dispersion statistics, thresholds, decision order and error definitions are given in Supplementary Figs. S21–S25 and the Supplementary Methods.

Data availability

All physical seeds, fixed field hashes, serialized prompts, raw model outputs, parsed actions, trajectories and analysis tables are available in the code repository, <https://github.com/tkEzaki/observation-maps-llm-agents>, and are archived on Zenodo at <https://doi.org/10.5281/zenodo.21834781> (concept DOI; version v1.0.0 corresponds to <https://doi.org/10.5281/zenodo.21834782>). The complete acquisition traces, one record per model call for all 15 acquisition groups, are included; no separate data archive has to be obtained. Access to proprietary provider services is not required to reproduce the deterministic engine and statistical analyses, but exact re-acquisition may depend on continued availability of the specific provider model IDs and backend behaviour.

Code availability

Code for the engine, encoders, acquisition runners, statistical analyses and figure generation is available at <https://github.com/tkEzaki/observation-maps-llm-agents>, together with the hash manifest, the fixed model identifiers and generation parameters, and the per-call records of the response contract.

Use of generative artificial intelligence in manuscript preparation

Generative language models were used to assist with language editing of the manuscript, and with drafting and refactoring parts of the simulation, analysis and figure-generation code. All scientific claims, numerical values, statistical procedures, analyses and final text were reviewed and verified by the authors, who take full responsibility for them. Within the research pipeline itself, language models appear only as the experimental subject: they generate the agent actions that constitute the data. No language-model output was used as an automated rule to select, filter, impute or statistically adjudicate results; all scientific interpretations and decisions were made and verified by the authors.

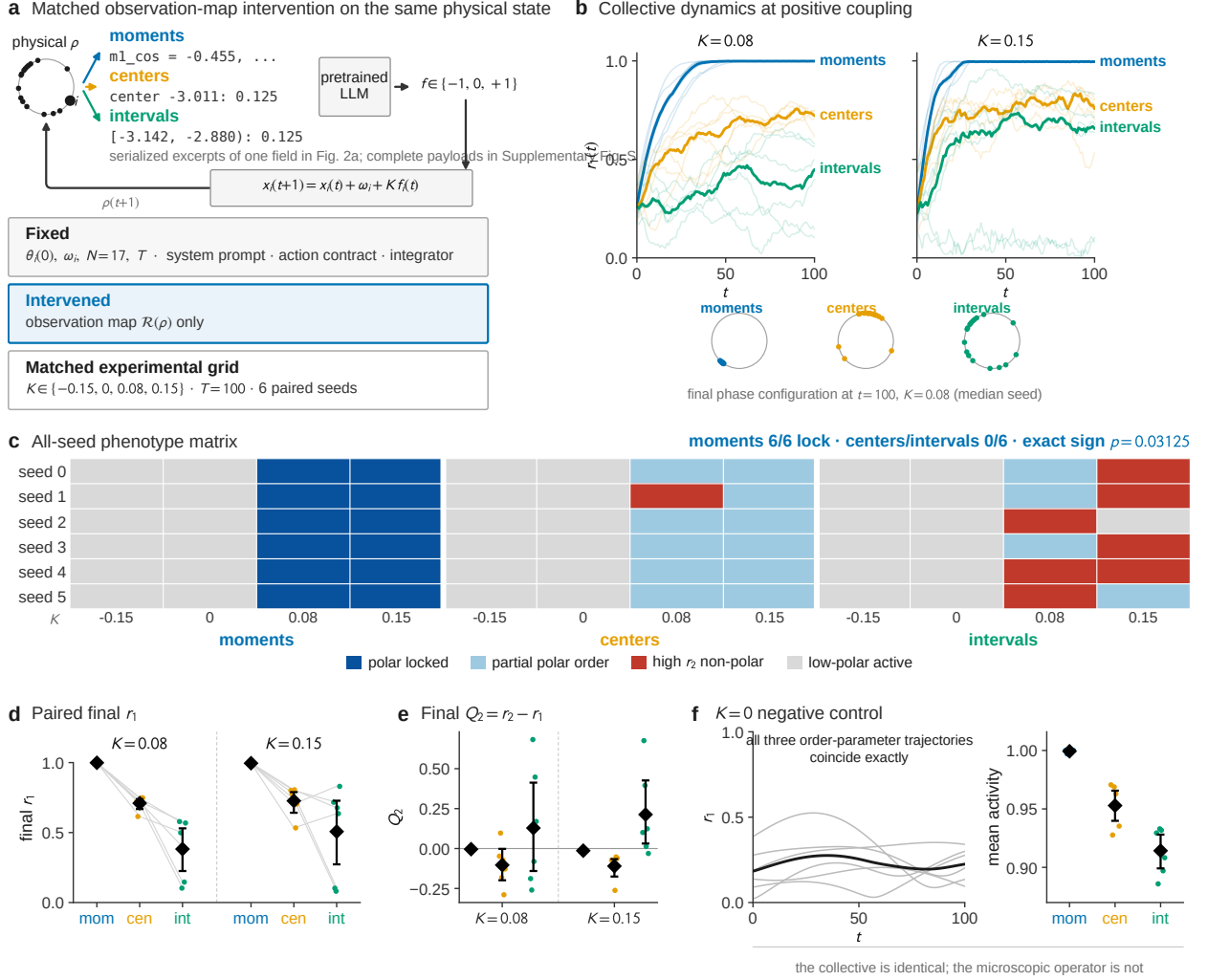


Figure 1: Observation maps select distinct collective outcomes in matched GPT agents. **a**, Matched observation-map intervention on the same physical state. One physical state (enlarged dot: the focal agent) is encoded through the three observation maps; the language model returns $f_i \in \{-1, 0, +1\}$ and the deterministic engine applies the coupling. Everything but the encoding was matched, including the instruction apart from its one-line observation description. **b**, Polar-order trajectories $r_1(t)$ at $K = 0.08$ and 0.15 ; thin lines are physical seeds and thick lines summarize the six seeds. **c**, Operational phenotype for every seed and condition. At each positive coupling, moments locked in 6/6 seeds and both histogram encodings in 0/6. **d**, Seed-level final r_1 at positive coupling. **e**, Relative second-harmonic order $Q_2 = r_2 - r_1$; intervals retained higher second-harmonic order relative to polar order. **f**, At $K = 0$, order-parameter trajectories coincide exactly across maps despite encoding-dependent activity. The physical seed, not the call, is the inference unit. Supporting evidence: engine audit (Supplementary Fig. S1), complete observation maps and prompt contracts (S2), every trajectory (S3), the phenotype classification under alternative thresholds (S4) and the negative-coupling condition (S5).

a Same physical field, three serialized observations

id 87f618

excerpts; complete serialisations in Supplementary Fig. S2

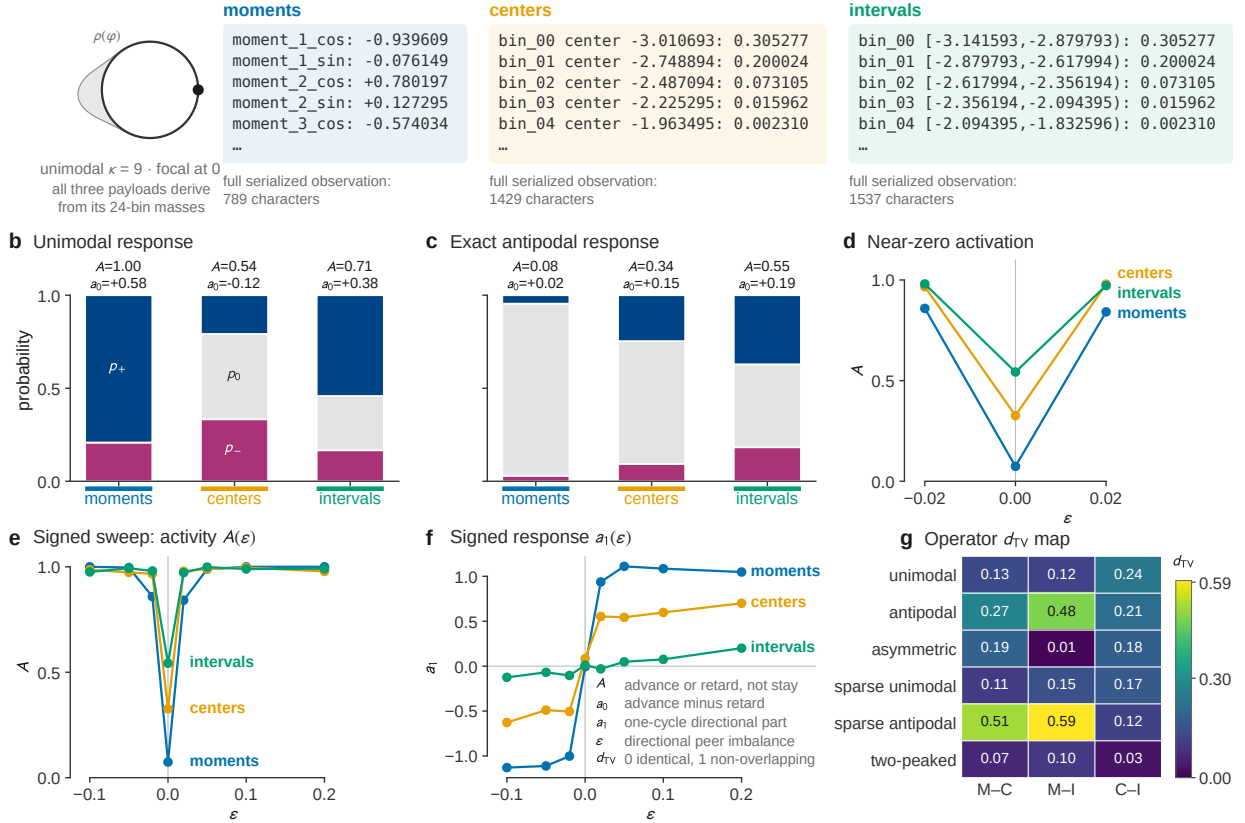


Figure 2: Controlled fields elicit encoding-dependent microscopic response operators. The three encodings turn the same field into three qualitatively different action rules. **a**, One unimodal relative-phase field, visualized by its generating von Mises density, and excerpts of the three serialized observations derived from its 24-bin masses (complete payloads in Supplementary Fig. S2). **b**, Full trinomial responses to the unimodal field. **c**, Responses to exact antipodal balance. **d**, Activity near zero signed imbalance. **e,f**, Activity and signed polar response across the imbalance grid. The sharp moments activation is resolution-limited and is not interpreted as a discontinuity. **g**, Pairwise total-variation distance between response operators across controlled field families. Repeated calls estimate the action probabilities; where uncertainty is shown, the resampling unit is specified for the corresponding analysis, and repeated calls to one input are not treated as independent experimental replicates. Supporting evidence: validation of the offset design (Supplementary Fig. S6), the preliminary eight-contract screen that motivated treating the encoding as an intervention (S7), and the complete sweeps behind **b** (S8), **c-f** (S9) and the sparse finite-peer fields (S10).

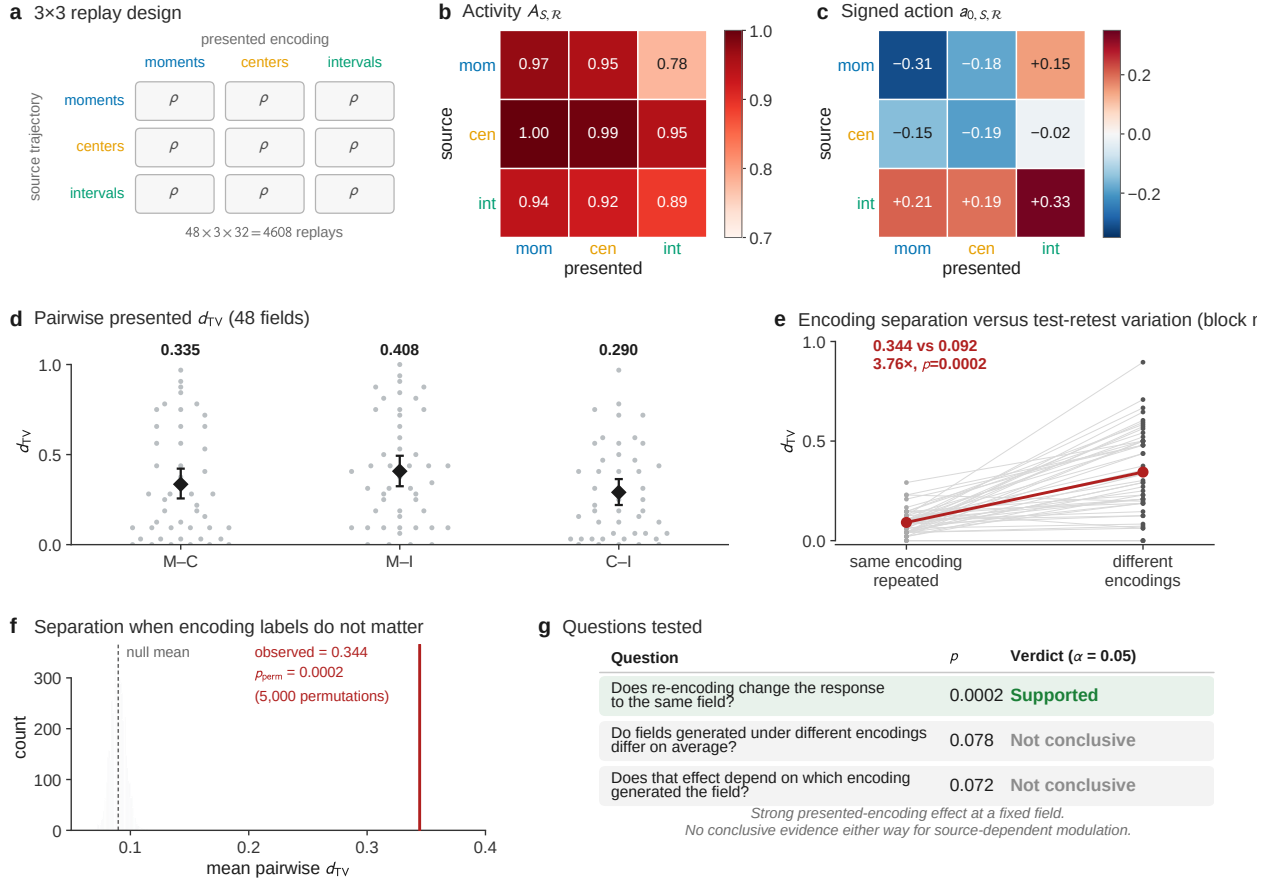


Figure 3: Identical-field replay isolates a presented-encoding effect. **a**, Forty-eight physical fields were fixed before the replay acquisition, balanced across source encodings and six trajectory strata, then re-encoded under all three presented maps. Each field–presented-encoding cell contained 32 responses in two blocks. Throughout **b,c,g**, rows are where the field came from (its source encoding) and columns are how that same frozen field was shown during replay (its presented encoding). **b**, Mean activity by source and presented encoding. **c**, Mean signed action by source and presented encoding. **d**, Field-level pairwise presented-encoding total-variation distances with field-cluster 95% bootstrap confidence intervals. **e**, Between-presented-encoding separation compared with the within-presented-encoding between-block noise floor for the same physical fields. **f**, Field-blocked presented-encoding-label permutation distribution; observed mean pairwise $d_{TV} = 0.344$, $p = 0.0002$. **g**, The three prespecified tests. The presented-encoding effect was supported; neither test for field dependence crossed the prespecified threshold, and neither establishes its absence (source main effect $p = 0.078$; source-by-presented-encoding interaction $p = 0.072$; Supplementary Fig. S13e–h). The 48 physical fields, not the 4,608 calls, are the inference units. Supporting evidence: the field-selection rule and its audit, which make the panel independent of the outcome it tests (Supplementary Fig. S11), every field’s three response distributions unpooled (S12), and the permutation nulls and stratified decomposition (S13).

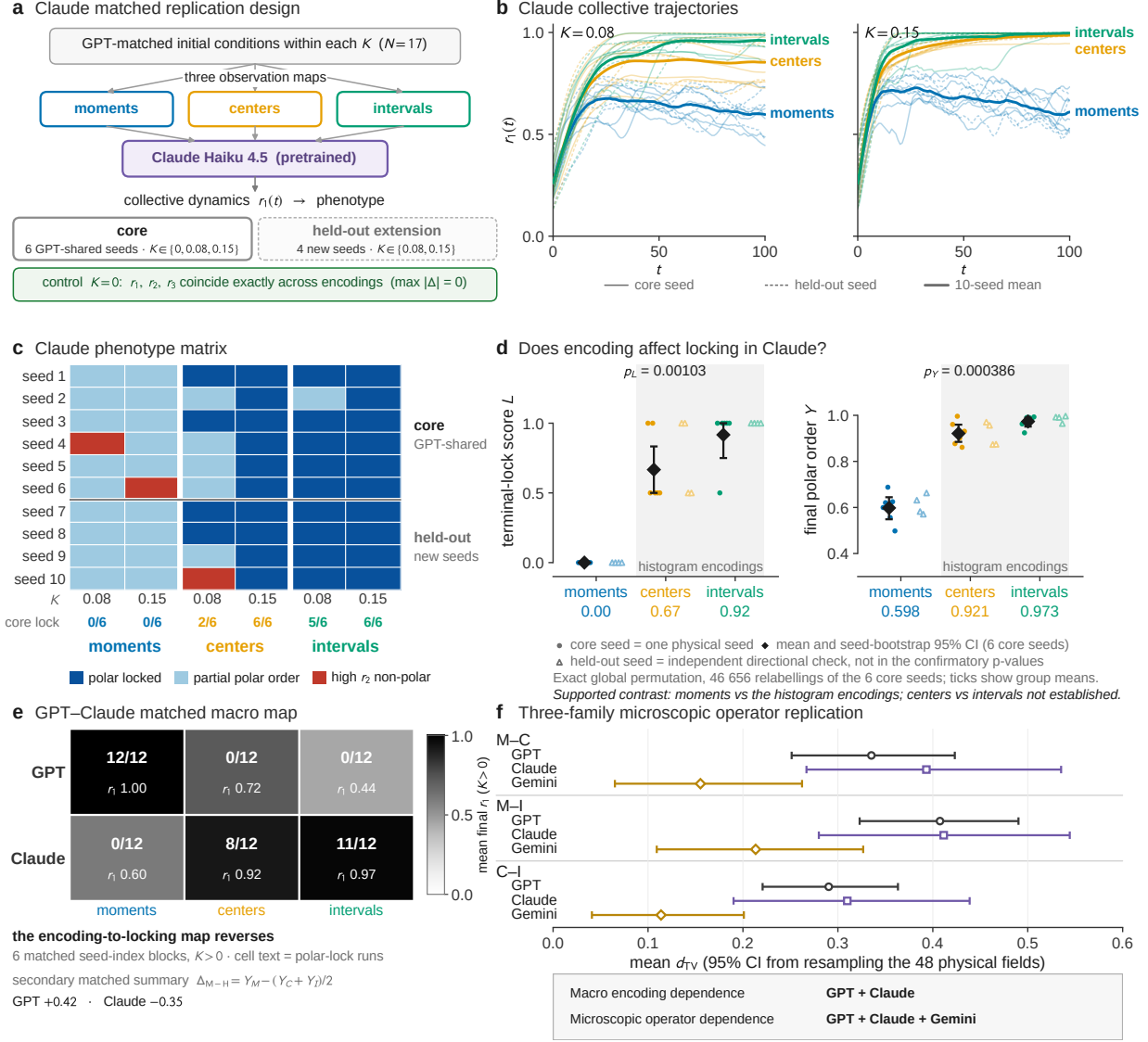
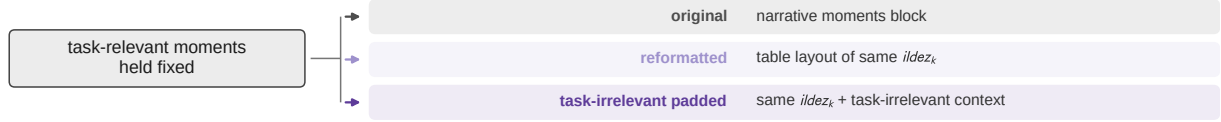
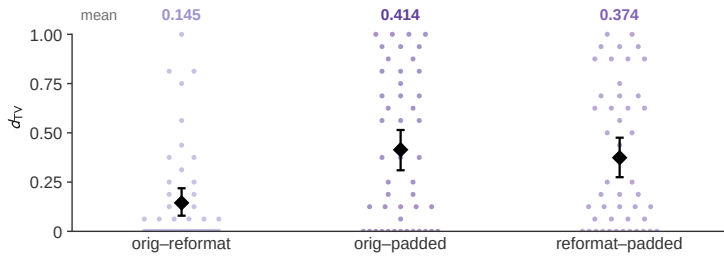


Figure 4: Observation maps select model-family-specific collective outcomes. **a**, Claude matched macro design: six seed-index blocks matched to GPT within each coupling value at $K = 0, 0.08, 0.15$, plus four held-out seeds at positive coupling. **b**, Claude $r_1(t)$ trajectories. At $K = 0.15$, centers and intervals lock while moments remains partially ordered. **c**, Claude phenotype matrix, separating confirmatory and held-out seeds. **d**, Prespecified core inference. Mean terminal-lock scores were 0.00, 0.67, 0.92 and mean final-order scores were 0.598, 0.921, 0.973 for moments, centers and intervals, respectively. Exact within-seed global permutations gave $p = 0.00103$ and $p = 0.000386$. **e**, GPT–Claude matched macro map using the six matched seed-index blocks. The observation-map effect replicated in Claude, but the lock/nonlock map reversed. **f**, Identical-field microscopic operator separation in GPT, Claude and Gemini. Macro dependence was tested in GPT and Claude; Gemini macro dynamics were not tested. Supporting evidence: the complete three-family operator comparison (Supplementary Fig. S14), every Claude trajectory (S15), the exact within-seed inference and the prespecified replication criteria (S16), and the seed-by-seed reversal (S17).

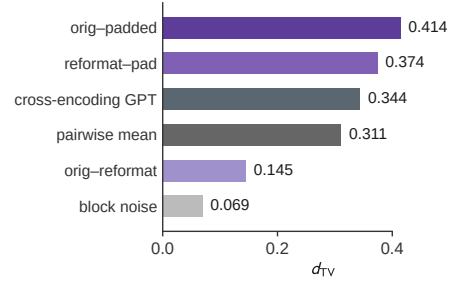
a Same task-information control: identical ($\bar{z}_1, \bar{z}_2, \bar{z}_3$)



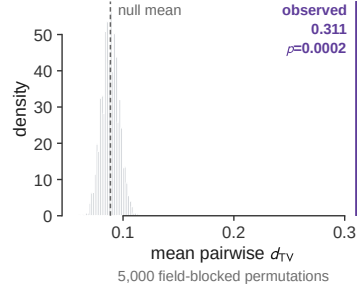
b Pairwise same-task-info d_{TV}



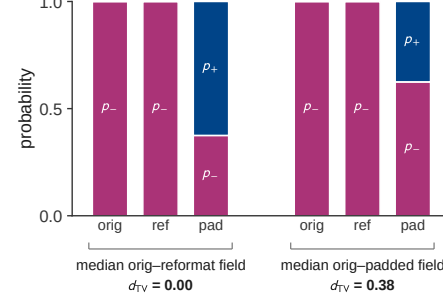
c Scale comparison



d If presentation labels did not matter



e Typical response patterns



reformatting leaves the median field unchanged; padding does not

f Effective observation map $\mathcal{R}(\rho)$

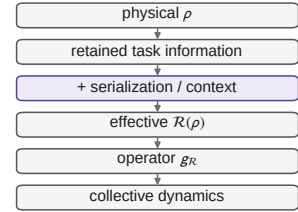


Figure 5: Presentation changes alter the GPT response operator when task-relevant numerical information is fixed. **a**, The same moment values presented in the original narrative, a reformatted table and a task-irrelevant-padding condition with added task-irrelevant context. **b**, Field-level pairwise total-variation distances. Means were 0.145 for original–reformatted, 0.414 for original–task-irrelevant-padding and 0.374 for reformatted–task-irrelevant-padding. **c**, Comparison with within-variant block noise (0.069) and the GPT cross-encoding reference (0.344). **d**, Field-blocked variant-label permutation; global mean pairwise $d_{TV} = 0.311$, $p = 0.0002$. **e**, Representative fields, each the field nearest the median of that contrast; the distance printed beneath is that contrast’s median over all 48 fields, not the selected field’s own. **f**, Retained task information and textual presentation jointly determine the effective observation map. The padding condition changed context, value position and length together and is not interpreted as a length-only effect. Supporting evidence: the complete control (Supplementary Fig. S18), and two further controls reported in the Supplementary Information only, varying length at fixed information (S19) and crossing the two textual binding features (S20).

References

1. Sclar, M., Choi, Y., Tsvetkov, Y. & Suhr, A. Quantifying language models' sensitivity to spurious features in prompt design, or: how I learned to start worrying about prompt formatting. In *Proc. International Conference on Learning Representations (ICLR)* (2024). Preprint at <https://arxiv.org/abs/2310.11324>.
2. Lu, Y., Bartolo, M., Moore, A., Riedel, S. & Stenetorp, P. Fantastically ordered prompts and where to find them: overcoming few-shot prompt order sensitivity. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL)* 8086–8098 (2022). <https://doi.org/10.18653/v1/2022.acl-long.556>.
3. Liu, N. F. et al. Lost in the middle: how language models use long contexts. *Trans. Assoc. Comput. Linguist.* **12**, 157–173 (2024). https://doi.org/10.1162/tac1_a_00638.
4. Zheng, C., Zhou, H., Meng, F., Zhou, J. & Huang, M. Large language models are not robust multiple choice selectors. In *Proc. International Conference on Learning Representations (ICLR)* (2024). Preprint at <https://arxiv.org/abs/2309.03882>.
5. Park, J. S. et al. Generative agents: interactive simulacra of human behavior. In *Proc. 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)* (ACM, 2023). <https://doi.org/10.1145/3586183.3606763>.
6. Li, H. et al. Theory of mind for multi-agent collaboration via large language models. In *Proc. 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 180–192 (2023). <https://doi.org/10.18653/v1/2023.emnlp-main.13>.
7. Du, Y., Li, S., Torralba, A., Tenenbaum, J. B. & Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *Proc. 41st International Conference on Machine Learning*, PMLR **235**, 11733–11763 (2024). Preprint at <https://arxiv.org/abs/2305.14325>.
8. Chen, W. et al. AgentVerse: facilitating multi-agent collaboration and exploring emergent behaviors. In *Proc. International Conference on Learning Representations (ICLR)* (2024). Preprint at <https://arxiv.org/abs/2308.10848>.
9. Ashery, A. F., Aiello, L. M. & Baronchelli, A. Emergent social conventions and collective bias in LLM populations. *Sci. Adv.* **11**, eadu9368 (2025). <https://doi.org/10.1126/sciadv.adu9368>.
10. Zomer, N. & De Domenico, M. Unraveling the emergence of collective behavior in networks of cognitive agents. *npj Artif. Intell.* **2**, 36 (2026). <https://doi.org/10.1038/s44387-026-00091-5>.
11. Willis, R., Zhao, J., Du, Y. & Leibo, J. Z. Evaluating collective behaviour of hundreds of LLM agents. Preprint at <https://arxiv.org/abs/2602.16662> (2026).
12. Ezaki, T., Imura, N. & Nishinari, K. When shippers become algorithms: candidate exposure, information design, and the concentration of LLM-mediated freight markets. Preprint at <https://arxiv.org/abs/2607.19967> (2026).
13. La Malfa, E. et al. Large language models miss the multi-agent mark. In *Advances in Neural Information Processing Systems* 38, Position Paper Track (2025). Preprint at <https://arxiv.org/abs/2505.21298>.
14. Perdomo, J. C., Zrnic, T., Mendler-Dünner, C. & Hardt, M. Performative prediction. In *Proc. 37th International Conference on Machine Learning*, PMLR **119**, 7599–7609 (2020). Preprint at <https://arxiv.org/abs/2002.06673>.
15. Ross, S., Gordon, G. J. & Bagnell, J. A. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proc. 14th International Conference on Artificial Intelligence and Statistics (AISTATS)* 627–635 (2011). Preprint at <https://arxiv.org/abs/1011.0686>.
16. Zhang, A., McAllister, R., Calandra, R., Gal, Y. & Levine, S. Learning invariant representations for reinforcement learning without reconstruction. In *Proc. International Conference on Learning Representations (ICLR)* (2021). Preprint at <https://arxiv.org/abs/2006.10742>.

17. Kaelbling, L. P., Littman, M. L. & Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artif. Intell.* **101**, 99–134 (1998). [https://doi.org/10.1016/S0004-3702\(98\)00023-X](https://doi.org/10.1016/S0004-3702(98)00023-X).
18. Kuramoto, Y. Self-entrainment of a population of coupled non-linear oscillators. In *International Symposium on Mathematical Problems in Theoretical Physics*, Lecture Notes in Physics vol. 39, 420–422 (Springer, 1975). <https://doi.org/10.1007/BFb0013365>.
19. Winfree, A. T. Biological rhythms and the behavior of populations of coupled oscillators. *J. Theor. Biol.* **16**, 15–42 (1967). [https://doi.org/10.1016/0022-5193\(67\)90051-3](https://doi.org/10.1016/0022-5193(67)90051-3).
20. Acebrón, J. A., Bonilla, L. L., Pérez Vicente, C. J., Ritort, F. & Spigler, R. The Kuramoto model: a simple paradigm for synchronization phenomena. *Rev. Mod. Phys.* **77**, 137–185 (2005). <https://doi.org/10.1103/RevModPhys.77.137>.
21. Kiss, I. Z., Zhai, Y. & Hudson, J. L. Emerging coherence in a population of chemical oscillators. *Science* **296**, 1676–1678 (2002). <https://doi.org/10.1126/science.1070757>.
22. Dörfler, F., Chertkov, M. & Bullo, F. Synchronization in complex oscillator networks and smart grids. *Proc. Natl Acad. Sci. USA* **110**, 2005–2010 (2013). <https://doi.org/10.1073/pnas.1212134110>.
23. Yamaguchi, S. et al. Synchronization of cellular clocks in the suprachiasmatic nucleus. *Science* **302**, 1408–1412 (2003). <https://doi.org/10.1126/science.1089287>.
24. Hansel, D., Mato, G. & Meunier, C. Clustering and slow switching in globally coupled phase oscillators. *Phys. Rev. E* **48**, 3470–3477 (1993). <https://doi.org/10.1103/PhysRevE.48.3470>.
25. Okuda, K. Variety and generality of clustering in globally coupled oscillators. *Physica D* **63**, 424–436 (1993). [https://doi.org/10.1016/0167-2789\(93\)90121-G](https://doi.org/10.1016/0167-2789(93)90121-G).
26. Daido, H. Order function and macroscopic mutual entrainment in uniformly coupled limit-cycle oscillators. *Prog. Theor. Phys.* **88**, 1213–1218 (1992). <https://doi.org/10.1143/ptp/88.6.1213>.
27. Kiss, I. Z., Zhai, Y. & Hudson, J. L. Predicting mutual entrainment of oscillators with experiment-based phase models. *Phys. Rev. Lett.* **94**, 248301 (2005). <https://doi.org/10.1103/PhysRevLett.94.248301>.
28. Xie, C. et al. Can large language model agents simulate human trust behavior? In *Advances in Neural Information Processing Systems* 37 (2024). Preprint at <https://arxiv.org/abs/2402.04559>.
29. Cemri, M. et al. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems* 38, Datasets and Benchmarks Track (2025). Preprint at <https://arxiv.org/abs/2503.13657>.
30. Leiding, A., van Rooij, R. & Shutova, E. The language of prompting: what linguistic properties make a prompt successful? In *Findings of the Association for Computational Linguistics: EMNLP 2023* 9210–9232 (2023). <https://doi.org/10.18653/v1/2023.findings-emnlp.618>.
31. Min, S. et al. Rethinking the role of demonstrations: what makes in-context learning work? In *Proc. 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 11048–11064 (2022). <https://doi.org/10.18653/v1/2022.emnlp-main.759>.
32. Givan, R., Dean, T. & Greig, M. Equivalence notions and model minimization in Markov decision processes. *Artif. Intell.* **147**, 163–223 (2003). [https://doi.org/10.1016/S0004-3702\(02\)00376-4](https://doi.org/10.1016/S0004-3702(02)00376-4).
33. Lesort, T., Díaz-Rodríguez, N., Goudou, J.-F. & Filliat, D. State representation learning for control: an overview. *Neural Netw.* **108**, 379–392 (2018). <https://doi.org/10.1016/j.neunet.2018.07.006>.
34. Fish, S., Gonczarowski, Y. A. & Shorrer, R. I. Algorithmic collusion by large language models. Preprint at <https://arxiv.org/abs/2404.00806> (2024).
35. Calvano, E., Calzolari, G., Denicolò, V. & Pastorello, S. Artificial intelligence, algorithmic pricing, and collusion. *Am. Econ. Rev.* **110**, 3267–3297 (2020). <https://doi.org/10.1257/aer.20190623>.

36. Li, N., Gao, C., Li, M., Li, Y. & Liao, Q. EconAgent: large language model-empowered agents for simulating macroeconomic activities. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics (ACL)* 15523–15536 (2024). <https://doi.org/10.18653/v1/2024.acl-long.829>.
37. Ha, D. & Schmidhuber, J. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems* 31 (2018). Preprint at <https://arxiv.org/abs/1803.10122>.
38. Hafner, D., Lillicrap, T., Ba, J. & Norouzi, M. Dream to control: learning behaviors by latent imagination. In *Proc. International Conference on Learning Representations (ICLR)* (2020). Preprint at <https://arxiv.org/abs/1912.01603>.
39. Janner, M., Fu, J., Zhang, M. & Levine, S. When to trust your model: model-based policy optimization. In *Advances in Neural Information Processing Systems* 32 (2019). Preprint at <https://arxiv.org/abs/1906.08253>.
40. Hardt, M. & Mendler-Dünner, C. Performative prediction: past and future. *Statist. Sci.* **40**, 417–436 (2025). <https://doi.org/10.1214/25-ST5986>.
41. Narang, A., Faulkner, E., Drusvyatskiy, D., Fazel, M. & Ratliff, L. J. Multiplayer performative prediction: learning in decision-dependent games. *J. Mach. Learn. Res.* **24**(202), 1–56 (2023). Preprint at <https://arxiv.org/abs/2201.03398>.
42. Kelly, M., Sidrane, C., Driggs-Campbell, K. & Kochenderfer, M. J. HG-Dagger: interactive imitation learning with human experts. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)* 8077–8083 (2019). <https://doi.org/10.1109/ICRA.2019.8793698>.
43. Koh, P. W. et al. WILDS: a benchmark of in-the-wild distribution shifts. In *Proc. 38th International Conference on Machine Learning*, PMLR **139**, 5637–5664 (2021). Preprint at <https://arxiv.org/abs/2012.07421>.
44. Strogatz, S. H. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D* **143**, 1–20 (2000). [https://doi.org/10.1016/S0167-2789\(00\)00094-4](https://doi.org/10.1016/S0167-2789(00)00094-4).
45. Chen, L., Zaharia, M. & Zou, J. How is ChatGPT’s behavior changing over time? *Harvard Data Sci. Rev.* **6**(2) (2024). <https://doi.org/10.1162/99608f92.5317da47>.
46. Wu, Q. et al. AutoGen: enabling next-gen LLM applications via multi-agent conversations. In *Proc. Conference on Language Modeling (COLM)* (2024). Preprint at <https://arxiv.org/abs/2308.08155>.
47. Wang, J., Wang, J., Athiwaratkun, B., Zhang, C. & Zou, J. Mixture-of-agents enhances large language model capabilities. In *Proc. International Conference on Learning Representations (ICLR)* (2025). Preprint at <https://arxiv.org/abs/2406.04692>.

Author contributions

T.E.: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, writing (original draft), writing (review and editing). N.I.: supervision, writing (review and editing). K.N.: supervision, writing (review and editing).

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests

The authors declare no competing interests.

Correspondence

Correspondence and requests for materials should be addressed to T.E.

Supplementary Information

Contents

Supplementary Information	26
Claim-to-evidence map	28
Analysis guide: what each statistical operation does	28
Supplementary Figure S1: Deterministic engine validation and causal isolation	30
Supplementary Figure S2: Complete observation maps, prompt contracts and serialization metadata	31
Supplementary Figure S3: Complete GPT matched-collective trajectories	33
Supplementary Figure S4: Robustness of the GPT collective phenotype classification	36
Supplementary Figure S5: Negative coupling and nonlocking dynamical pathways	37
Supplementary Figure S6: Prespecified offset design and aliasing control for response harmonics	38
Supplementary Figure S7: Complete encoding-invariance screen	39
Supplementary Figure S8: Full concentration-dependent microscopic transmutation map	40
Supplementary Figure S9: Antipodal symmetry and dense signed-imbalance response	41
Supplementary Figure S10: Sparse finite-peer fields and variance decomposition	42
Supplementary Figure S11: Prespecified replay-field selection and audit	43
Supplementary Figure S12: Complete fieldwise replay response panel	44
Supplementary Figure S13: Replay robustness, source and interaction tests	48
Supplementary Figure S14: Complete three-family microscopic operator replication	49
Supplementary Figure S15: Complete Claude matched-collective trajectories	50
Supplementary Figure S16: Claude controls, exact inference and prespecified gates	53
Supplementary Figure S17: GPT–Claude reversal of the encoding-to-phenotype map	54
Supplementary Figure S18: Complete same-task-information observation-map control	55
Supplementary Figure S19: Serialization-length control	57
Supplementary Figure S20: Serialization-binding control	59
Supplementary Figure S21: Surrogate analysis: what is being tested	61
Supplementary Figure S22: In-distribution compressibility does not guarantee closed-loop support	62

Supplementary Figure S23: Moments surrogate development, replay diagnosis and prospective repair	63
Supplementary Figure S24: Centers (in-distribution compressibility without closed-loop support)	65
Supplementary Figure S25: Intervals (matching descriptors did not yield consistent responses)	66
Supplementary Table S1: Serialized observation lengths and provider token counts	68
Supplementary Table S2: Acquisition parameters and response-parser disposition	69
Supplementary Methods: Fourier fitting conventions	69
Supplementary Methods: surrogate compressibility and transportability	70
Reproducibility and decision audit	74

Claim-to-evidence map

The map below lets a reader locate, for each main claim of the study, the main-text figure that states it and the Supplementary Figures that carry the supporting evidence.

Main claim	Main figure	Supplementary Figures
Exact causal isolation ($K = 0$ negative control)	Fig. 1	S1, S2
GPT collective outcomes	Fig. 1	S3–S5
Controlled microscopic transmutation	Fig. 2	S6–S10
Identical-field operator effect	Fig. 3	S11–S13
Three-family microscopic replication	Fig. 4	S14
Claude macro reversal	Fig. 4	S15–S17
Same-task-information control	Fig. 5	S18
Serialization length versus retained information	Supplementary Information only	S19
Which serialization feature is involved	Supplementary Information only	S20
Surrogate compressibility and transportability	Supplementary Information only	S21–S25

Analysis guide: what each statistical operation does

To preclude any ambiguity about what was computed, the table below states what each statistical operation used in the main text and in this Supplementary Information does.

Term	What it does
Permutation test	Swaps the labels between conditions many times and asks how often chance alone produces a difference as large as the observed one.
Exact sign test	Counts only the direction of the paired differences, not their size.
Bootstrap confidence interval	Resamples the unit stated for that analysis in the note below this table, and reports the range of the recomputed estimate.
Total-variation distance	The difference between two advance/stay/retard probability distributions, on a scale from 0 (identical) to 1 (no overlap).
Block variation (“block noise”)	The test-retest difference obtained when the same condition is measured again in a separate acquisition batch.
Omnibus test	Asks whether any of several groups differ, without identifying which pair does.
Permutation p floor	With 5,000 resamples the observed statistic is counted among them (plus-one), so the smallest reportable p is $1/5001 \approx 0.0002$. A p at the floor means “none of 5,000 relabellings was as extreme”, not a measured zero.
Out-of-fold prediction	Predicts a held-out fold using a model fitted without it. What is placed in the held-out fold decides what is being tested, so the holdout scheme is named wherever an out-of-fold number is reported.
Support	Whether the training set contains fields similar to the test field, and whether those similar fields elicited consistent actions.

Four levels of surrogate evaluation

The surrogate analysis (Supplementary Figs. S21–S25) asks four increasingly demanding questions, and they are not interchangeable. Naming them separately avoids the common conflation in which passing a cross-validation split is reported as evidence that a surrogate can be deployed inside a running simulation.

1. *Cross-block stability*. Does the response law measured on a field in one acquisition block reproduce on a second, independent block of the *same* field? This is the `acquisition_block_holdout` scheme, and it is the criterion by which the production model class was selected. It is a reproducibility check: the same field appears in training and in test.
2. *In-domain field generalization*, referred to as *compressibility*. Can the surrogate predict controlled fields that were excluded from training altogether? This is what the four field-preserving schemes measure (leave-one-profile-out, offset-group, sparse-realization and source-family holdout).
3. *Closed-loop support*. Do the fields an interacting population generates for itself fall inside the region the training set covers, and do their near neighbours in that set agree with one another about what the model does?
4. *Prospective closed-loop accuracy*, referred to as *transportability*. Does the surrogate predict new language-model collective runs it played no part in producing?

We use *transportability* only for level 4, and “grouped out-of-fold, by field” refers to level 2 only. Failure at an earlier level stopped a branch before the later, more expensive level was attempted: only moments reached level 4.

Names used in the released code

In legacy artifact names and analysis tables, *representation* denotes the observation map of the main text and *target* denotes the presented encoding; source encoding is *source* in both.

Figure numbers and released file names

The supplementary figures were reordered into narrative order after the analysis was frozen, and the released scripts and PDFs keep their original numbers. The complete mapping between manuscript figure numbers and released file names is provided in the repository README.

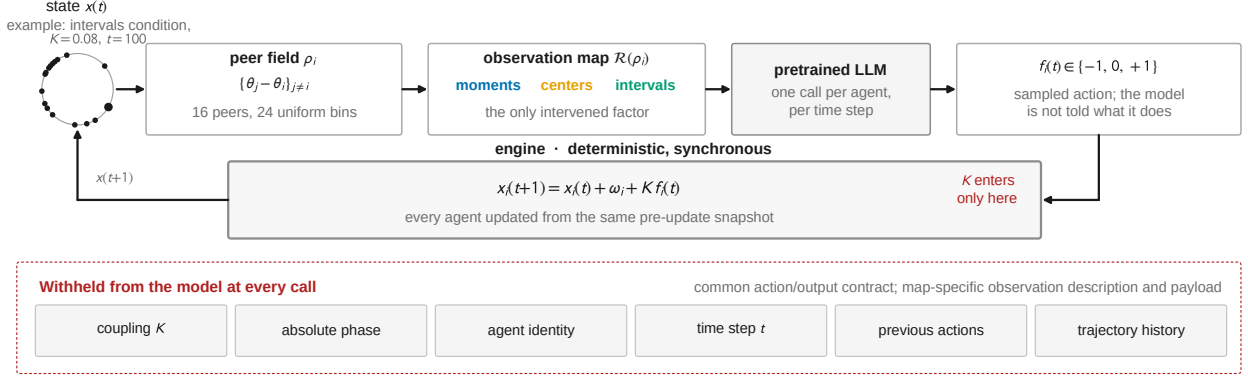
Multiplicity across the three replay tests

The three global replay tests (presented encoding, source, source $\times$ presented encoding) are reported without multiplicity correction, each carrying its own prespecified $\alpha = 0.05$, because they answer separate prespecified questions rather than dividing one. The reported decisions do not depend on that choice: a Bonferroni correction across the three would leave the presented-encoding effect supported ($p = 0.0002$ against $\alpha/3 = 0.0167$) and would leave the source and interaction tests non-significant either way.

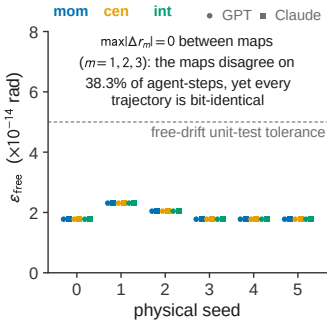
The unit that is shuffled or resampled differs between experiments and is stated with each analysis: permutations are carried out within a field or within a seed, and fields and seeds are never mixed with one another.

Throughout the captions, panels marked descriptive or secondary are post hoc and not used for inference, and all conclusions are limited to the tested maps, descriptors and acquisition domains.

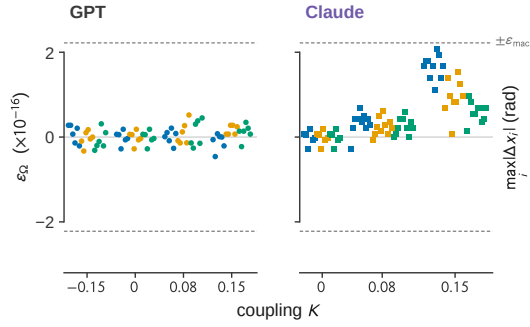
a Engine and information-flow contract



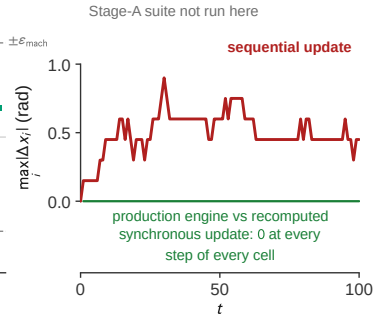
b Exact $K=0$ free evolution



c Collective-frequency accounting identity



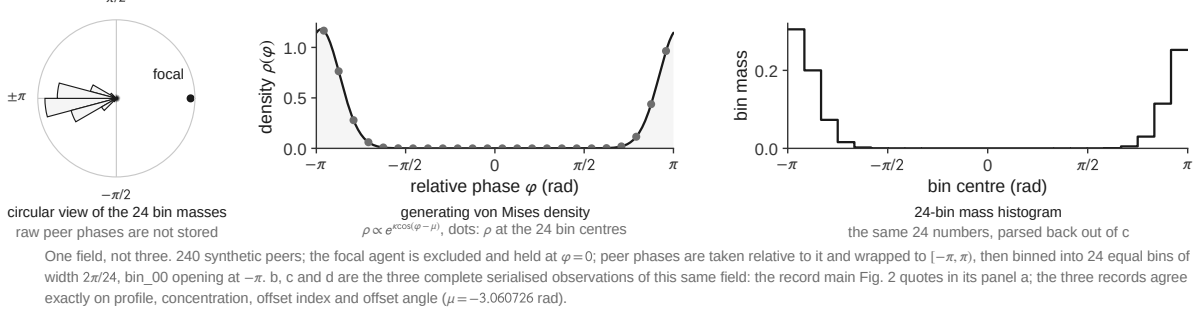
d Synchronous-update audit



Supplementary Figure S1. Deterministic engine validation and causal isolation. **a**, Engine and information-flow contract. The three maps share one action/output contract and differ only in the observation description and payload. The relative-phase multiset $\{\theta_j - \theta_i\}_{j \neq i}$ is reduced to the field ρ_i , encoded by $\mathcal{R}$ and passed to the language model, which returns one action $f_i \in \{-1, 0, +1\}$. All agents update synchronously and the engine alone applies $K f_i$ in $x_i(t+1) = x_i(t) + \omega_i + K f_i(t)$. Information never given to the model is marked: K , absolute phase, agent identity, time step, previous actions and history. The circle shows a final configuration (GPT intervals, $K = 0.08, t = 100$), the enlarged dot marking focal agent $i = 0$. **b**, Exact $K = 0$ free evolution. The maximum residual $\epsilon_{\text{free}} = \max_{i,t} |x_i(t) - x_i(0) - t\omega_i|$ over all 36 $K = 0$ cells (2 families $\times$ 3 maps $\times$ 6 seeds), one point per physical seed and observation map. The coupling term vanishes, so the sampled actions are behaviourally inert; the maximum, 2.31×10^{-14} rad, is floating-point accumulation. **c**, Collective-frequency accounting identity. The residual $\epsilon_{\Omega} = \Omega_{\text{coll}} - (\bar{\omega} + K\tau_{\text{social}})$ across all GPT and Claude collective runs, coloured by encoding; $\max|\epsilon_{\Omega}| = 2.08 \times 10^{-16}$. **d**, Synchronous-update audit. A deterministic toy trajectory at $K = 0.15$ under a peer-mass policy is run with the correct synchronous rule and with an incorrect sequential rule, and the production engine is checked against an independently recomputed synchronous update over all 150 cells. Scope: this figure validates the implementation, not the language-model response operator; encoder rotation and peer-permutation invariance were verified at the serialization level only.

a Stored example physical field: one physical field, three observation maps

unimodal $\kappa = 9 \cdot \text{offset index } 0 \cdot 240 \text{ peers}$



Prompt contract

lines 1 and 3–9 are byte-identical in b, c and d; only lines 2 and 10 differ, so all three are given. The user message is exactly these ten lines followed by the map's payload column below.

```

1 You are an interaction operator for a phase on a circle.
2 moments The values are the first three normalized circular moments of the distribution of other phases relative to you.
2 centers The table gives normalized relative-phase mass at 24 circular bin centers in radians.
2 intervals The table gives the normalized distribution of other phases relative to you in radians.
3 Negative relative phase is behind you; positive relative phase is ahead of you.
4 Choose one social correction in response to this distribution:
5 - "advance": move in the positive phase direction
6 - "stay": apply no social correction
7 - "retard": move in the negative phase direction
8 Use only the relative distribution below. Return exactly one JSON object and no other text.
9 Required schema: {"social_action": "advance|stay|retard"}
10 moments Relative phase distribution as circular moments:
10 centers Relative phase distribution by bin center:
10 intervals Relative phase distribution:

```

b Full moments observation

789 characters

```

moment_1_cos: -0.939609
moment_1_sin: -0.076149
moment_2_cos: +0.780197
moment_2_sin: +0.127295
moment_3_cos: -0.574034
moment_3_sin: -0.142058

```

lines 11–16, 6 numeric values. Harmonics $m = 1, 2, 3$ in order; within each harmonic the real (cos) component precedes the imaginary (sin) one. Signed, six decimals, fixed width. No per-bin mass is sent.

c Full centers observation

1429 characters

```

bin_00 center -3.010693: 0.305277
bin_01 center -2.748894: 0.200024
bin_02 center -2.487094: 0.073105
bin_03 center -2.225295: 0.015962
bin_04 center -1.963495: 0.002310
bin_05 center -1.701696: 0.000253
bin_06 center -1.439897: 0.000024
bin_07 center -1.178097: 0.000002
bin_08 center -0.916298: 0.000000
bin_09 center -0.654498: 0.000000
bin_10 center -0.392699: 0.000000
bin_11 center -0.130900: 0.000000
bin_12 center +0.130900: 0.000000
bin_13 center +0.392699: 0.000000
bin_14 center +0.654498: 0.000000
bin_15 center +0.916298: 0.000000
bin_16 center +1.178097: 0.000001
bin_17 center +1.439897: 0.000006
bin_18 center +1.701696: 0.000060
bin_19 center +1.963495: 0.000603
bin_20 center +2.225295: 0.005036
bin_21 center +2.487094: 0.030167
bin_22 center +2.748894: 0.114665
bin_23 center +3.010693: 0.252505

```

lines 11–34, 48 numeric values. 24 bin centres ascending from $-\pi + \pi/24$, standard origin (no half-bin shift). Centre signed, mass unsigned, both six decimals. 8 of the 24 bins print 0.000000 (grey): mass below 5×10^{-7} , not an empty bin.

d Full intervals observation

1537 characters

```

bin_00 [-3.141593, -2.879793): 0.305277
bin_01 [-2.879793, -2.617994): 0.200024
bin_02 [-2.617994, -2.356194): 0.073105
bin_03 [-2.356194, -2.094395): 0.015962
bin_04 [-2.094395, -1.832596): 0.002310
bin_05 [-1.832596, -1.570796): 0.000253
bin_06 [-1.570796, -1.308997): 0.000024
bin_07 [-1.308997, -1.047198): 0.000002
bin_08 [-1.047198, -0.785398): 0.000000
bin_09 [-0.785398, -0.523599): 0.000000
bin_10 [-0.523599, -0.261799): 0.000000
bin_11 [-0.261799, +0.000000): 0.000000
bin_12 [+0.000000, +0.261799): 0.000000
bin_13 [+0.261799, +0.523599): 0.000000
bin_14 [+0.523599, +0.785398): 0.000000
bin_15 [+0.785398, +1.047198): 0.000000
bin_16 [+1.047198, +1.308997): 0.000001
bin_17 [+1.308997, +1.570796): 0.000006
bin_18 [+1.570796, +1.832596): 0.000060
bin_19 [+1.832596, +2.094395): 0.000603
bin_20 [+2.094395, +2.356194): 0.005036
bin_21 [+2.356194, +2.617994): 0.030167
bin_22 [+2.617994, +2.879793): 0.114665
bin_23 [+2.879793, +3.141593): 0.252505

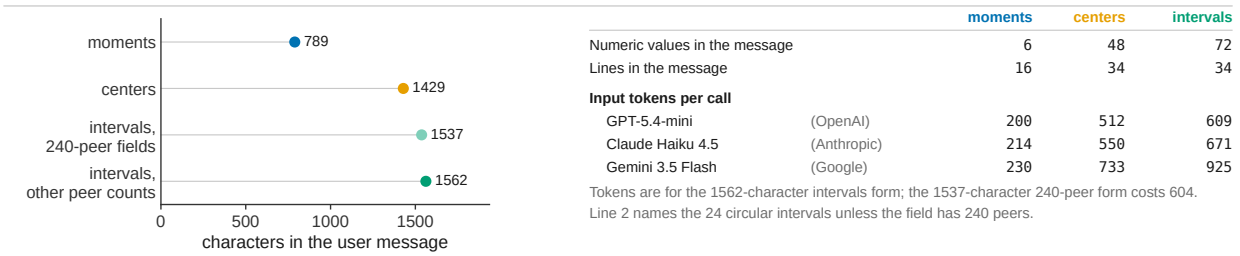
```

lines 11–34, 72 numeric values. Half-open [lo, hi): the lower edge belongs to the bin, the upper edge does not. bin_00 (bold) opens at $-\pi$; bin_23 (bold) closes at $+\pi$, which wraps back onto $-\pi$. 8 of the 24 bins print 0.000000 (grey): mass below 5×10^{-7} , not an empty bin.

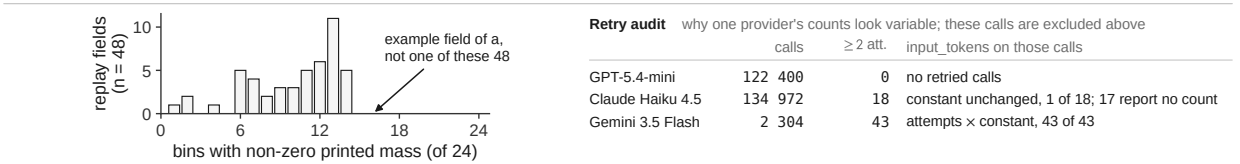
Supplementary Figure S2 (page 1 of 2). Complete observation maps, prompt contracts and serialization metadata (the three complete payloads of one fixed field). **a**, The fixed example physical field (the same record shown in main Fig. 2: profile `unimodal_k9`, offset index 0, 240 peers, focal agent excluded and held at $\varphi = 0$, wrapping to $[-\pi, \pi)$ with 24 bins of width $2\pi/24$), drawn as three derived views: the serialized 24-bin masses on the circle, the generating von Mises density, and the 24-bin mass histogram parsed back out of the centers payload. Raw peer phases were not stored and none are synthesised; the moments agent received only the fixed moment features shown in panel b. **b–d**, The complete production-form user messages generated from this one physical record under the moments (b), centers (c) and intervals (d) observation maps. Each payload is a pure function of the 24-bin histogram drawn in a, so the text shown is what the model received; boundary conventions, bin origin, ordering and the fixed-width six-decimal formatting are annotated on the panels. Concatenating the prompt contract, using each map's own lines 2 and 10, with that map's payload column below reproduces the user message.

e Message length and field occupancy are different quantities

Representation-level constants one value per map, set by the fixed-width $\pm d$.dddddd serialiser and the tokenizer. Point values, not distributions.



Field-dependent quantity how sparse the physical field is, counted in bins, not characters. It does not change message length: both bin encodings print all 24 rows.



Across the 48 stored replay fields (16 peers each) 1–14 of the 24 bins print non-zero mass (median 11). The example field of panel a is denser: 16 of its 24 bins print non-zero mass, and the remaining 8 print 0.000000 only because their mass falls below 5×10^{-7} and rounds to zero at six decimals; every one of the 24 bins holds mass, none is empty. Occupancy is what the field controls; the message length above is not.

f Retained-feature inventory: the maps differ in content, not only in format

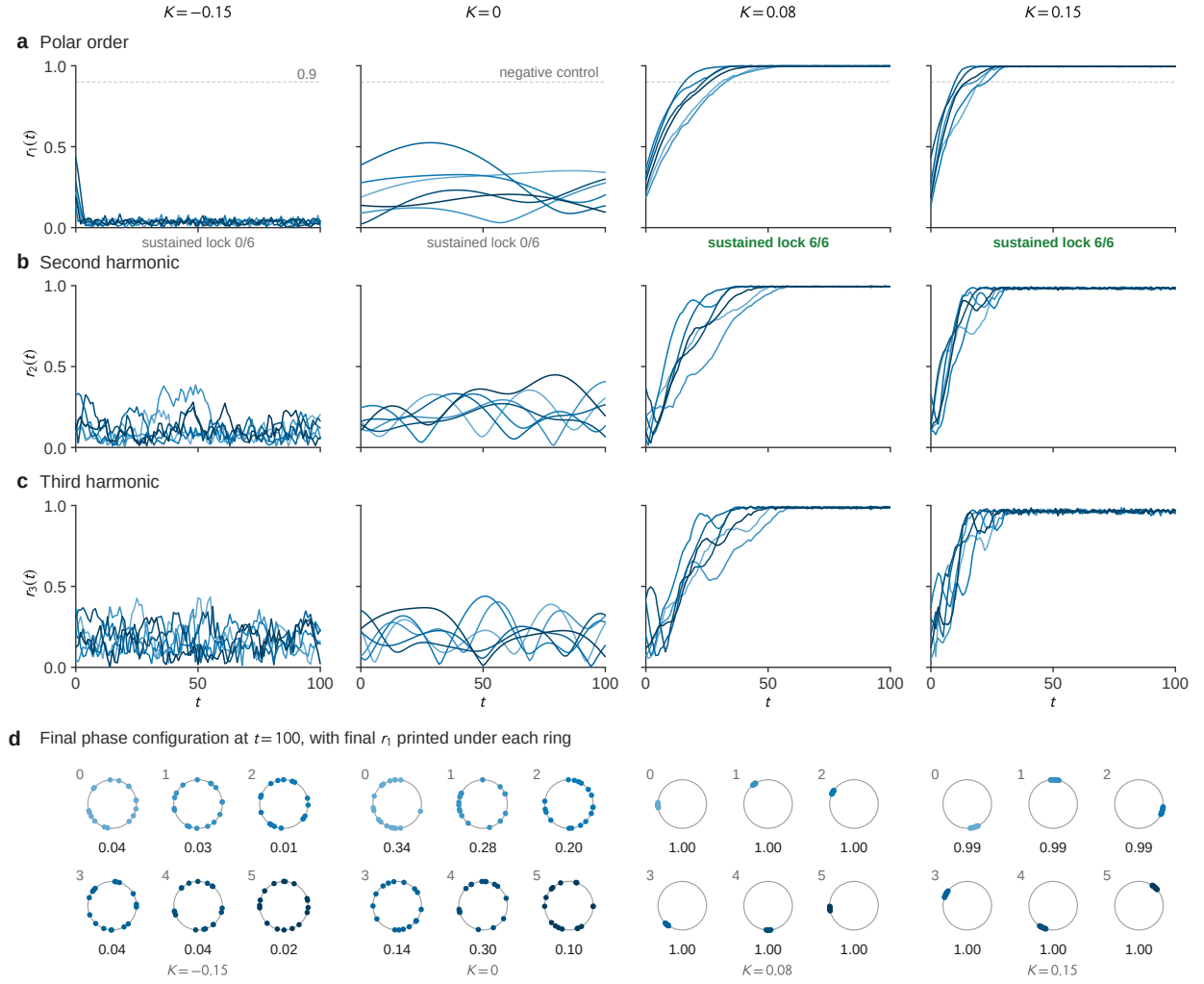
	moments	centers	intervals	what the difference does
First harmonic $m=1$ (polar order r_1, ψ_1)	✓	○	○	printed outright by moments; a 24-term sum for the other two
Second and third harmonics $m=2, 3$	✓	○	○	the same three complex numbers, arrived at differently
Harmonics $m \geq 4$	✗	○	○	discarded by moments: 24 masses carry 12 harmonics, moments keep 3
Per-bin mass, all 24 bins	✗	✓	✓	local detail; identical numbers in centers and intervals
Explicit bin-centre angle	✗	✓	✗	intervals leaves the centre implicit as the interval midpoint
Explicit half-open bin edges	✗	✗	✓	only intervals states the boundary convention in the text
Six-decimal fixed-width presentation	✓	✓	✓	identical numeric precision, so precision cannot explain the gap
Peer count / absolute phase / agent identity	✗	✗	✗	never serialised by any map
✓ printed explicitly	○ recoverable from what is printed		✗ not present in the message	

Supplementary Figure S2 (page 2 of 2). (continued: serialization metadata). e, Character and token counts. Because the serializers use fixed-width numeric formatting, character and token counts are encoding-level *constants*, not distributions, and are drawn as point values (Supplementary Table S1). The only field-dependent quantity is the number of occupied bins (1–14 of 24, median 11), a sparsity descriptor of the field shown on its own axis and explicitly not a proxy for prompt length. Retried calls are excluded from the token display. f, Retained-feature inventory: which physical information components (low-order polar information, higher-harmonic information, 24-bin local detail, explicit bin centers, explicit bin boundaries, six-decimal presentation) each map retains. The three maps are not pure formatting variants.

moments

observation map: every coupling, every seed, underlying main Fig. 1

seed 0 1 2 3 4 5

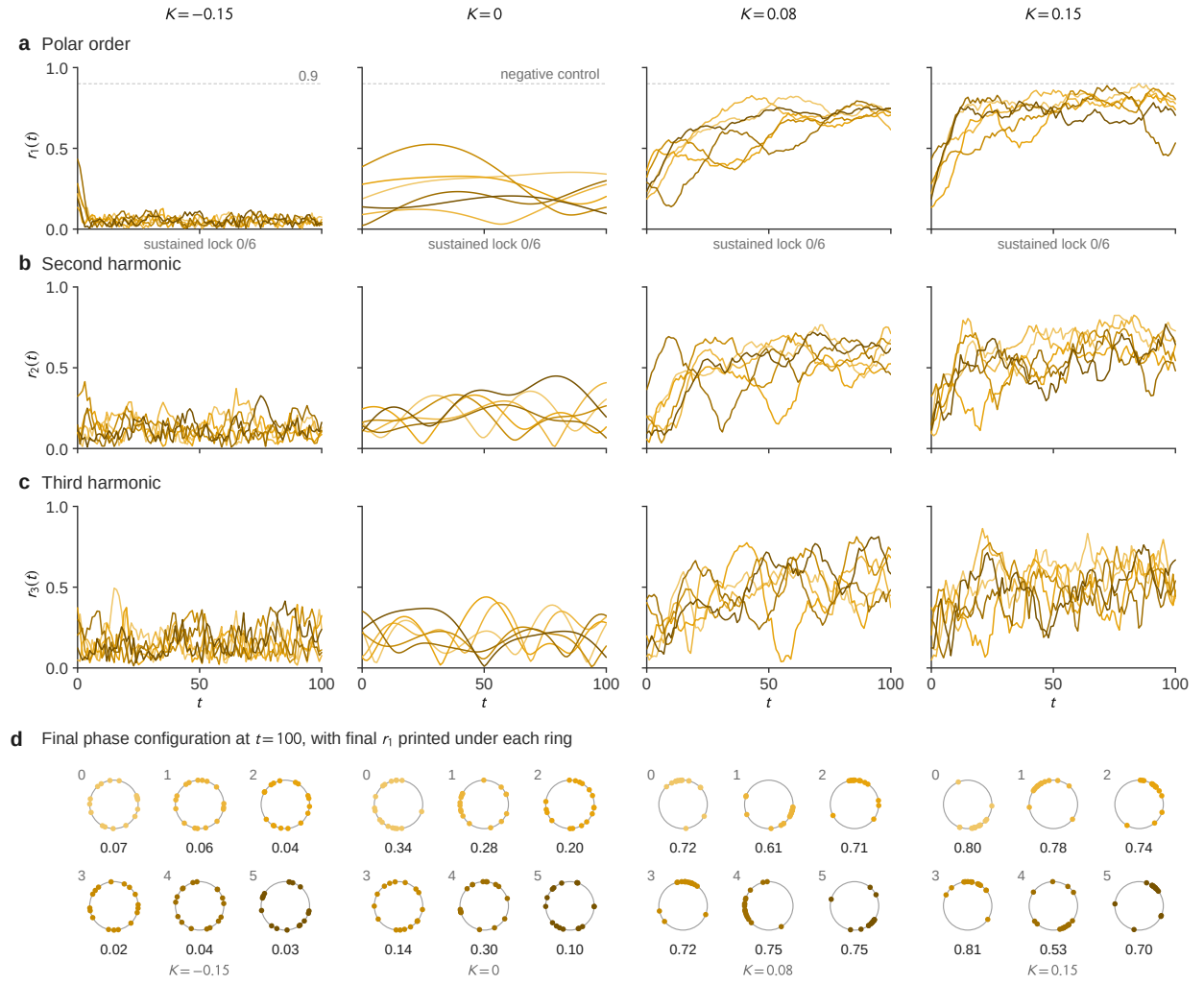
circular moments m_1, m_3 of the peer phases · $N=17$ oscillators, 100 steps · gpt-5.4-mini · 24 runs, 40,800 calls, 100.0% valid

Supplementary Figure S3 (page 1 of 3). Complete GPT matched-collective trajectories (one page per observation map; this page: moments). Complete trajectory sets underlying main Fig. 1, from the matched GPT experiment (gpt-5.4-mini, $N = 17$, $T = 100$, $K \in \{-0.15, 0, 0.08, 0.15\}$, six paired physical seeds, 122,400 valid calls). Each page shows, for one observation map, a 3×4 grid of (harmonic, coupling) cells with all six seed trajectories of $r_m(t)$ for $m = 1, 2, 3$ (nothing is averaged across seeds, so run-to-run timing variation remains visible), plus the 24 final phase configurations at $t = 100$. The y range of each harmonic is identical on all three pages. The dashed rule in **a** is the prespecified polar-locking threshold $r_1 = 0.9$. At $K = 0$ the order-parameter trajectories of the three maps coincide exactly (Supplementary Fig. S1b); at positive K the moments condition shows terminal lock in 6/6 seeds (the endpoint criterion $r_1(T) \geq 0.9$ gives the same counts).

centers

observation map: every coupling, every seed, underlying main Fig. 1

seed 0 1 2 3 4 5

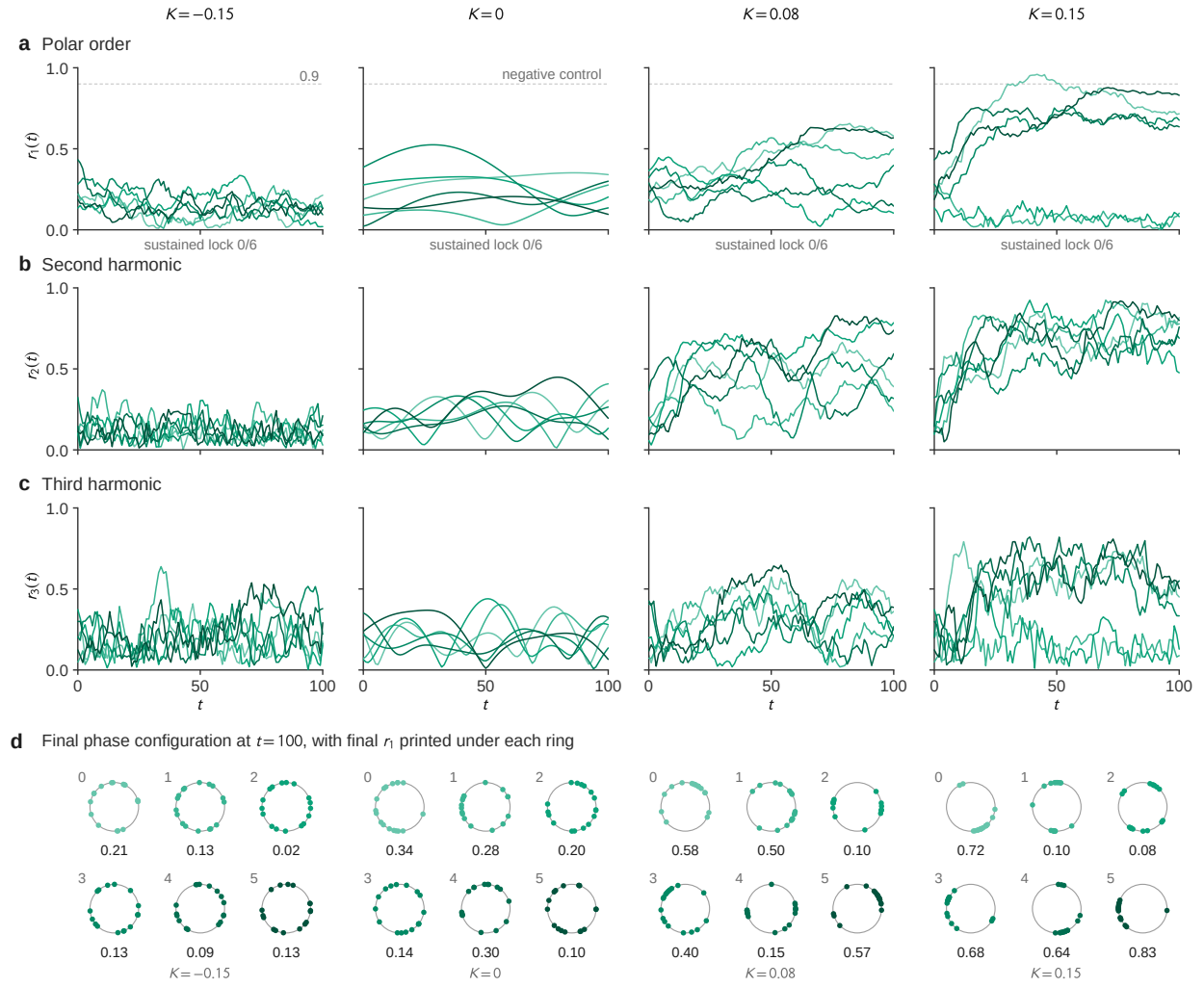
24-bin phase histogram, bin centres in standard notation · $N=17$ oscillators, 100 steps · gpt-5.4-mini · 24 runs, 40,800 calls, 100.0% valid

Supplementary Figure S3 (page 2 of 3). (continued: centers). As on page 1 of Supplementary Fig. S3, for the centers observation map, which locks in 0/6 seeds at each positive coupling and predominantly produces partial polar order.

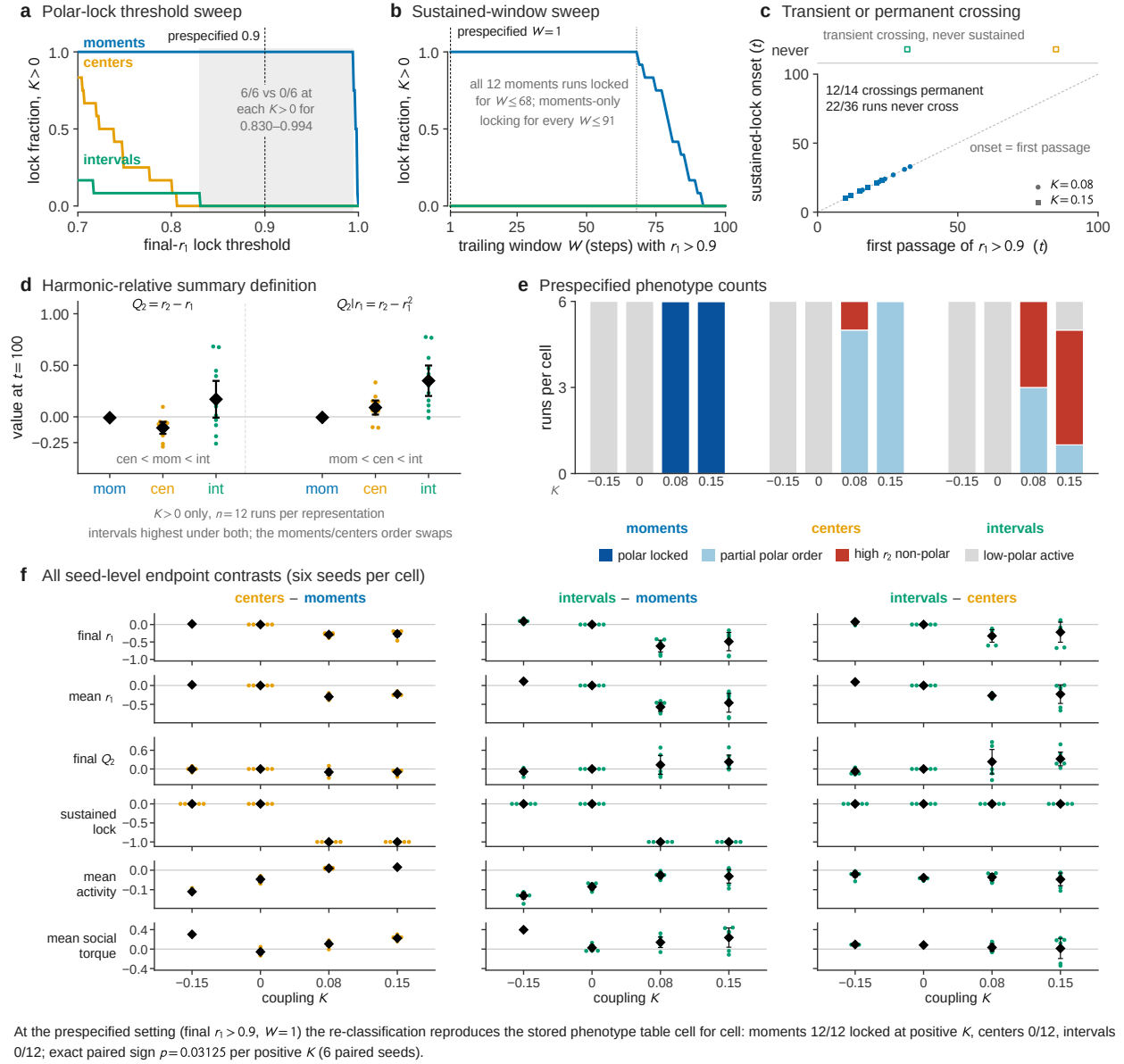
intervals

observation map: every coupling, every seed, underlying main Fig. 1

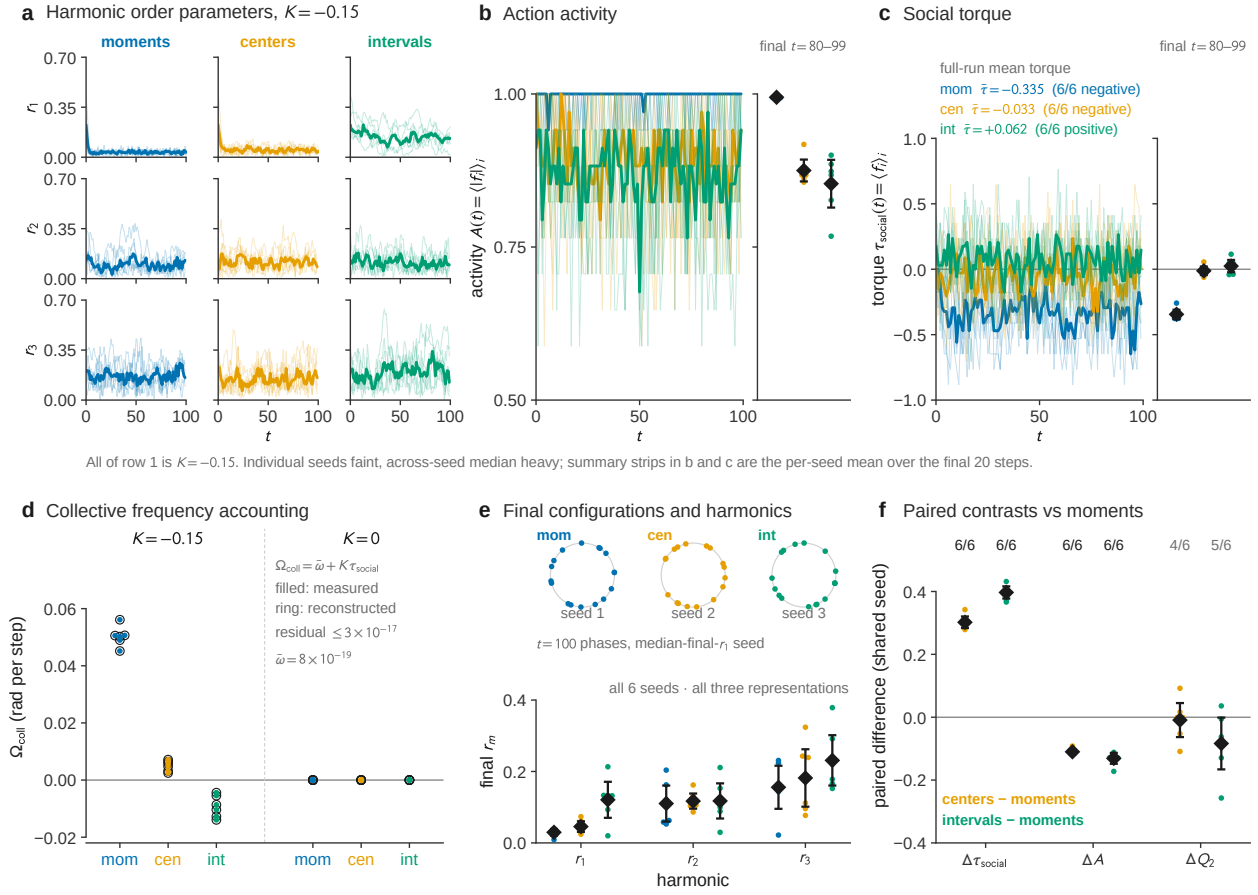
seed 0 1 2 3 4 5

24-bin phase histogram, bin intervals to six decimals · $N=17$ oscillators, 100 steps · gpt-5.4-mini · 24 runs, 40,800 calls, 100.0% valid

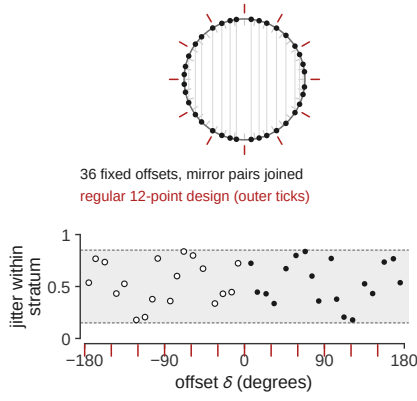
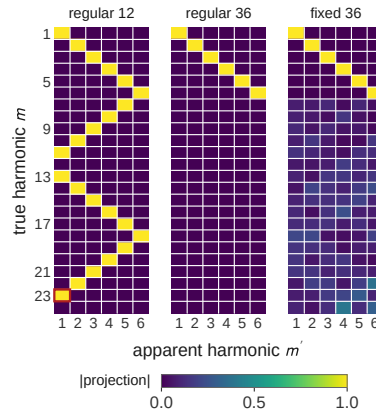
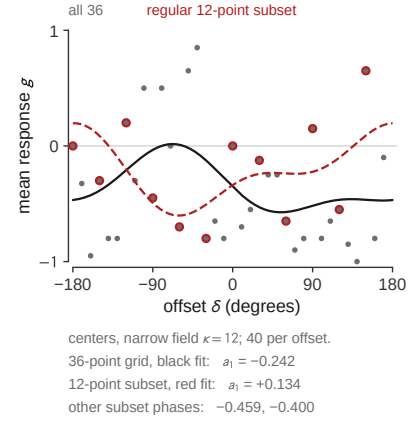
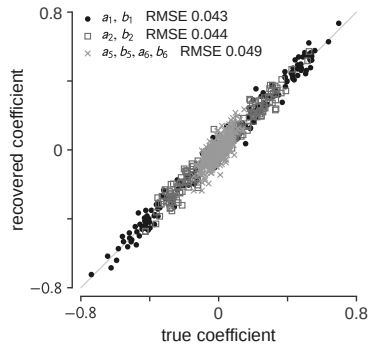
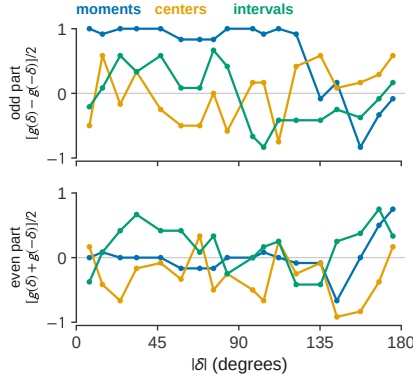
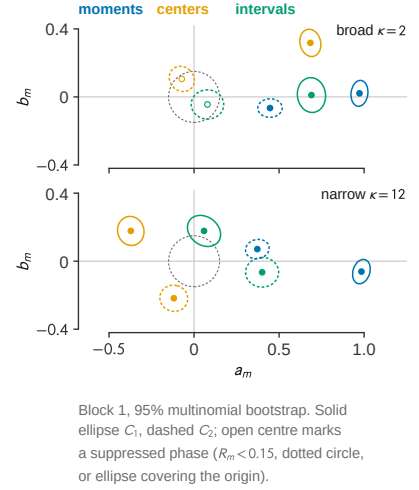
Supplementary Figure S3 (page 3 of 3). (continued: intervals). As on page 1 of Supplementary Fig. S3, for the intervals observation map, which locks in 0/6 seeds at each positive coupling and more often retains second-harmonic order relative to polar order ($Q_2 = r_2 - r_1$).



Supplementary Figure S4. Robustness of the GPT collective phenotype classification. Sensitivity of the operational headline (moments 6/6 locked, centers and intervals 0/6, exact paired sign test $p = 2/2^6 = 0.03125$) to the choices buried in the prespecified rule. Panels **a–d** re-classify the trajectories with the prespecified analysis functions. **a**, Lock fraction against the final- r_1 threshold, over a band around the prespecified 0.9. **b**, Lock fraction against a minimum duration: the length of the trailing window over which r_1 must remain above threshold. All 12 moments runs (six seeds at each positive coupling) remain locked for trailing windows of up to 68 steps, and for every tested window of up to 91 steps locking occurs only under moments, so the prespecified rule's admission of $t^* = T$ is not what produces the 6/6 versus 0/6 separation. **c**, First passage versus terminal-lock onset: whether each crossing of 0.9 is permanent or transient, run by run. **d**, The harmonic-relative summary $Q_2 = r_2 - r_1$ beside the one alternative already present in the artifacts, $r_2 - r_1^2$. **e**, Phenotype counts for every encoding and coupling. **f**, All seed-level endpoint contrasts (final r_1 , mean r_1 , final Q_2 , terminal lock, activity, torque), one row per endpoint and one column per encoding contrast. Scope: a sensitivity analysis around a prespecified rule, not a threshold search; no panel reports an optimised threshold, window or summary ($n = 72$ runs).



Supplementary Figure S5. Negative coupling and nonlocking dynamical pathways. The GPT $K = -0.15$ condition. **a**, All six seed trajectories of r_1, r_2, r_3 , faceted by encoding: polar order stays low for every map. **b**, Activity trajectories and final-window means: near-zero polar order coexists with near-unit activity, that is, almost no abstention. **c**, Social torque trajectories and final-window estimates; the full-run mean torques differ in sign and magnitude across maps (-0.335 moments, -0.033 centers, $+0.062$ intervals). **d**, Collective-frequency accounting: measured Ω_{coll} against $\bar{\omega} + K\tau_{\text{social}}$; the drift is coupling-mediated and vanishes at $K = 0$. **e**, Mechanically selected final phase configurations (one per encoding, not the maximum-difference seed) with all final order parameters. **f**, Shared-seed paired contrasts against moments in torque, activity and Q_2 ; counts give how many of the six shared seeds agree in sign. Torque and activity separate the encodings at every seed, Q_2 does not. Scope: the state is active polar-order suppression; panels **c** and **f** are not evidence of source-domain feedback.

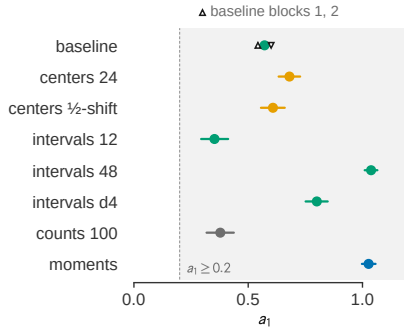
a Fixed nonuniform offset design**b** Aliasing onto low-order modes**c** Measured sign change**d** Recovery on measured offsets**e** Mirror-pair balance**f** Bootstrap stability of C_m 

Supplementary Figure S6. Prespecified offset design and aliasing control for response harmonics. The harmonic coefficients reported throughout the study are meaningful only if the sampled angles can separate the harmonics being fitted. **a**, The 36 prespecified nonuniform offsets (mirror-paired, stratified jitter) shown on the circle and on the linear $[-\pi, \pi]$ axis. **b**, Analytic aliasing map for a regular 12-point design. With evenly spaced angles, a fast harmonic can masquerade as a slower one when fitted (aliasing); the map shows how much of each unit-amplitude true harmonic would leak into the fitted low-order modes, highlighting the $m=23 \rightarrow m=1$ contamination the nonuniform design prevents. **c**, The same leak in practice: one measured response, from a separate acquisition sampled on a regular 36-point grid, fitted twice, once with all 36 points and once with only its regular 12-point first block; only the 12-point fit changes the fitted low-order coefficients. **d**, Ground-truth check on simulated data: harmonic mixtures with known coefficients, including an out-of-band $m=23$ nuisance component, are sampled on the measured offsets and refitted; recovered coefficients fall on the identity line. **e**, Mirror-pair balance: sums and differences of responses at $\pm\delta$, documenting odd/even decomposition stability. **f**, Bootstrap confidence ellipses for the complex kernels C_1 and C_2 of representative broad and narrow fields, with the prespecified phase-reporting gate (phase is reported only when $R_m \geq 0.15$ and the conditional bootstrap ellipse excludes the origin). Panels **a**, **b** and **d–f** use the block-1 design; **c** refits a separate regular 36-point grid acquisition and its 12-point first block.

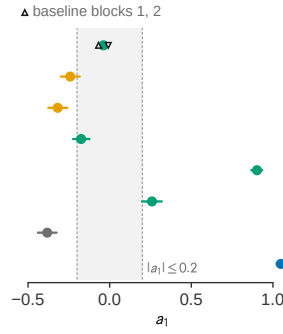
a Screened observation contracts

encoding	label in b–f	serialized example	feature encoded	bins	precision	value convention
24 intervals, 6 decimals (fixed baseline)	baseline		interval edges + mass	24	6 decimals	fractions, $\Sigma = 1$
24 bin centers	centers 24		bin center + mass	24	6 decimals	fractions, $\Sigma = 1$
24 bin centers, half-bin origin shift	centers ½-shift		bin center + mass	24	6 decimals	fractions, $\Sigma = 1$
12 intervals	intervals 12		interval edges + mass	12	6 decimals	fractions, $\Sigma = 1$
48 intervals	intervals 48		interval edges + mass	48	6 decimals	fractions, $\Sigma = 1$
24 intervals, 4 decimals	intervals d4		interval edges + mass	24	4 decimals	fractions, $\Sigma = 1$
24 integer counts, fixed total	counts 100		interval edges + count	24	6 decimals	integers, $\Sigma = 100$
circular moments	moments		cos/sin moments	n/a	6 decimals	$m = 1-3, \cos/\sin$

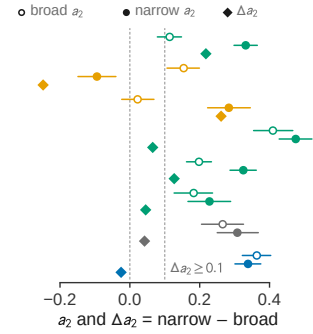
b Broad field $\kappa = 2$: first harmonic



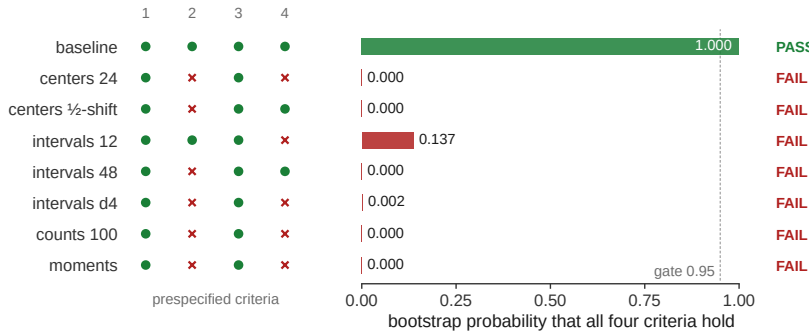
c Narrow field $\kappa = 12$: first harmonic



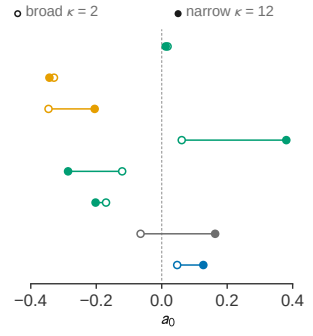
d Second harmonic



e Joint switching gate



f Mean action bias

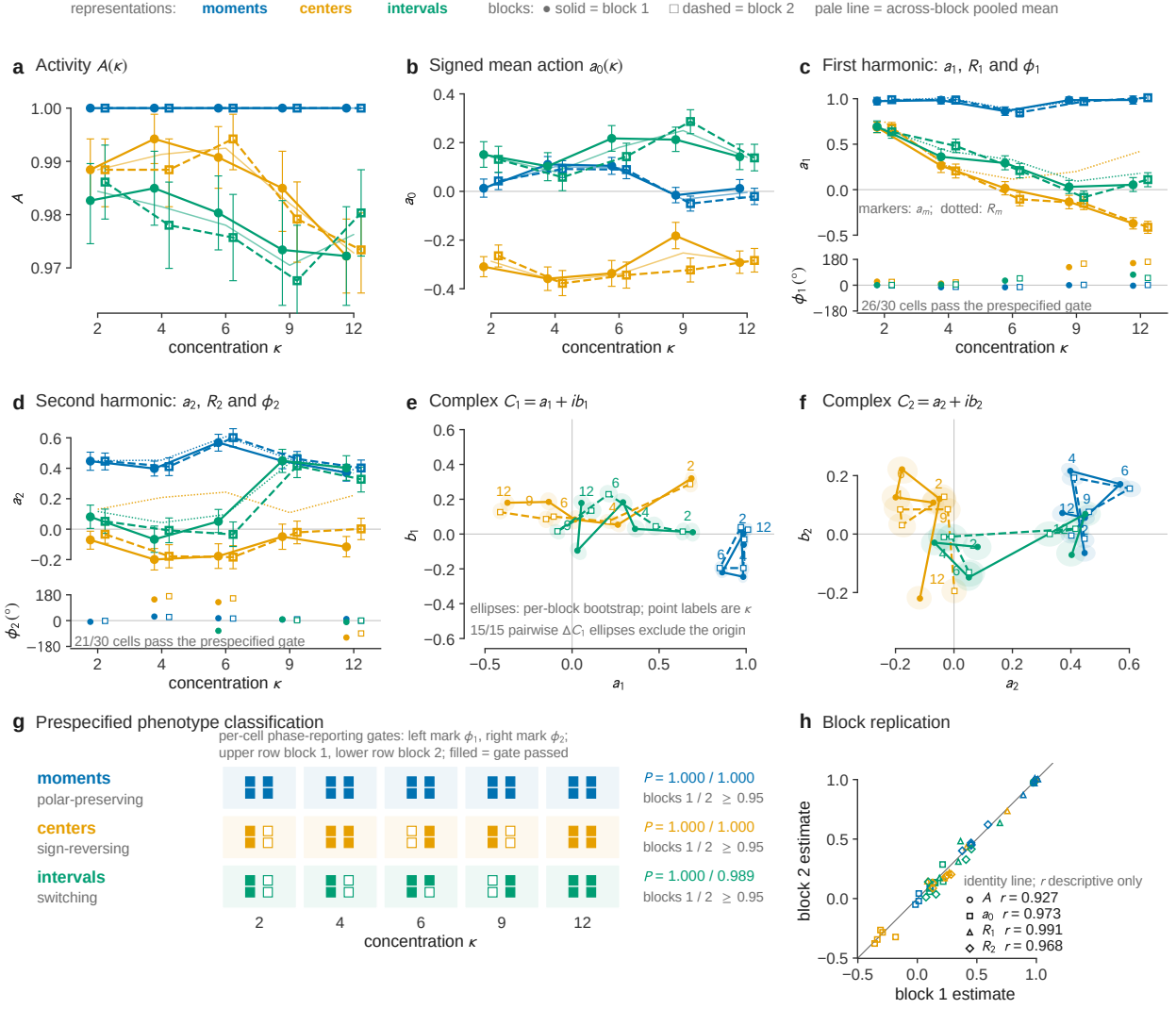


a Each example is the payload the fixed encoder actually emits for one common field ($\kappa = 2$, offset $\pi/4$), scaled to its own maximum.

b–f The seven alternates were acquired at 40 responses per condition on a 36-offset grid, the baseline at 100 on its 48-offset grid.

Points are the stored coefficients, bars bootstrap 95 % intervals; Δa_2 has no stored interval. e criteria: 1 broad $a_1 \geq 0.2$ · 2 narrow $|a_1| \leq 0.2$ · 3 broad $a_2 > 0$ · 4 $\Delta a_2 \geq 0.1$, on the point estimate.

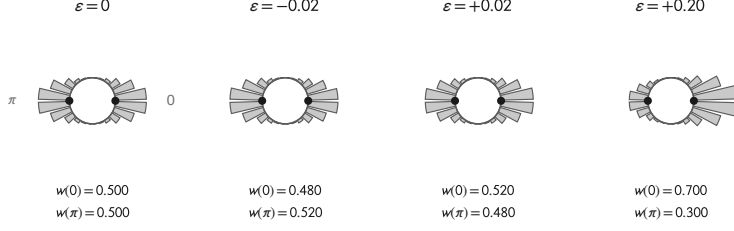
Supplementary Figure S7. Complete encoding-invariance screen. The preliminary screen that chose the three encodings used elsewhere, and the reason the observation map is treated as an intervention variable rather than a formatting detail. Eight observation contracts were compared on a common pair of fields before any collective experiment, asking whether the response operator survives changes a practitioner would call cosmetic. It did not. **a**, Screened encoding inventory (baseline intervals, centers, shifted bin origin, 12 bins, 48 bins, four-decimal precision, integer counts, moments); each example strip is the serialized payload for one common field. **b,c**, First-harmonic estimates a_1 with bootstrap intervals for the broad (**b**) and narrow (**c**) field, with the prespecified thresholds. **d**, Second harmonic a_2 for both fields and the switching contrast Δa_2 , with all prespecified gate thresholds. **e**, Joint probability of satisfying every prespecified switching criterion, per encoding, against the 0.95 gate, with the overall PASS/FAIL. **f**, Offset-averaged action bias a_0 per encoding: encoding changes shift average directional bias, not only harmonic shape. Scope: a motivating measurement, not a confirmatory result; no main-text claim rests on it. Only the baseline encoding was acquired in two blocks (open triangles in **b**, **c**), so no between-block strip exists for the seven alternates, and the three screened maps are not information-equivalent.



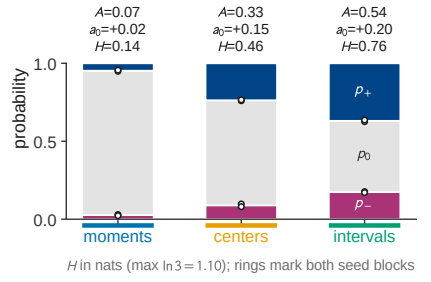
Supplementary Figure S8. Full concentration-dependent microscopic transmutation map. Complete operator map for one unimodal relative-phase field at five prespecified concentrations $\kappa \in \{2, 4, 6, 9, 12\}$, three observation maps, 36 offsets, 24 responses per condition and two independent acquisition blocks (25,920 valid responses at $M = 6$). **a**, Activity $A(\kappa)$ with block-specific estimates, pooled means and bootstrap intervals. **b**, Signed mean action $a_0(\kappa)$. **c,d**, First- and second-harmonic summaries: signed projections a_m , magnitudes $R_m = |C_m|$, and a phase strip showing ϕ_m only for cells passing the prespecified phase-reporting gate (Supplementary Fig. S6f). **e,f**, Complex-plane trajectories of C_1 (**e**) and C_2 (**f**) as κ increases, one connected path per encoding, with bootstrap ellipses. **g**, Prespecified microscopic phenotype classification per encoding and concentration (polar-preserving, switching, sign-reversing); all three rules passed in both blocks. The gate in **c,d** is a display rule for phase legibility; the phenotype rules are trajectory-level inference criteria reading the $\kappa = 2$ and $\kappa = 12$ endpoints only. **h**, Between-block replication: block-1 against block-2 estimates of A , a_0 , R_1 and R_2 on identity lines; trajectory correlations are $r(a_1) = 0.939\text{--}0.994$ and $r(a_2) = 0.868\text{--}0.986$, above the prespecified 0.8 threshold, and 39/45 block-difference intervals include zero. Error bars are 95% bootstrap intervals. Scope: block correlations, ellipses and origin exclusions are conditional on the two acquisition blocks.

a Antipodal field on the signed weight axis

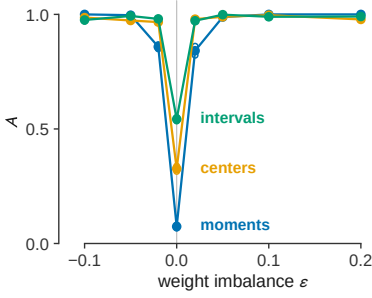
relative-phase mass, $\kappa = 6$; the ± 0.02 fields are not visually separable - serialisations in Supplementary Fig. S2



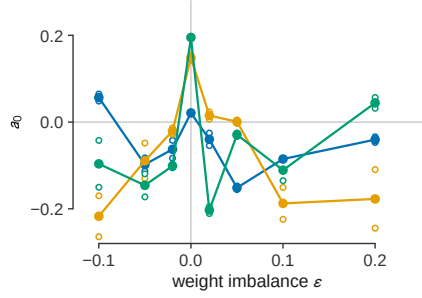
b Operator at exact balance



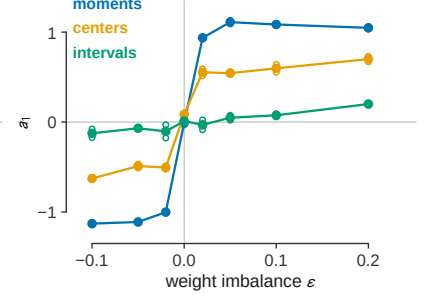
c Activity $A(\varepsilon)$



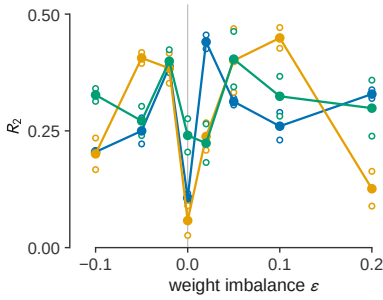
d Signed bias $a_0(\varepsilon)$



e Signed response $a_1(\varepsilon)$



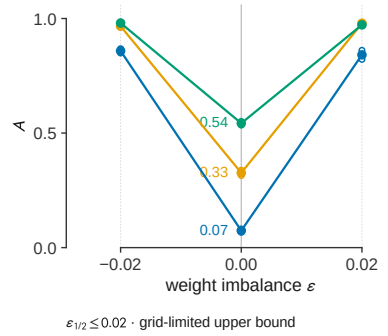
f Second harmonic $R_2(\varepsilon)$



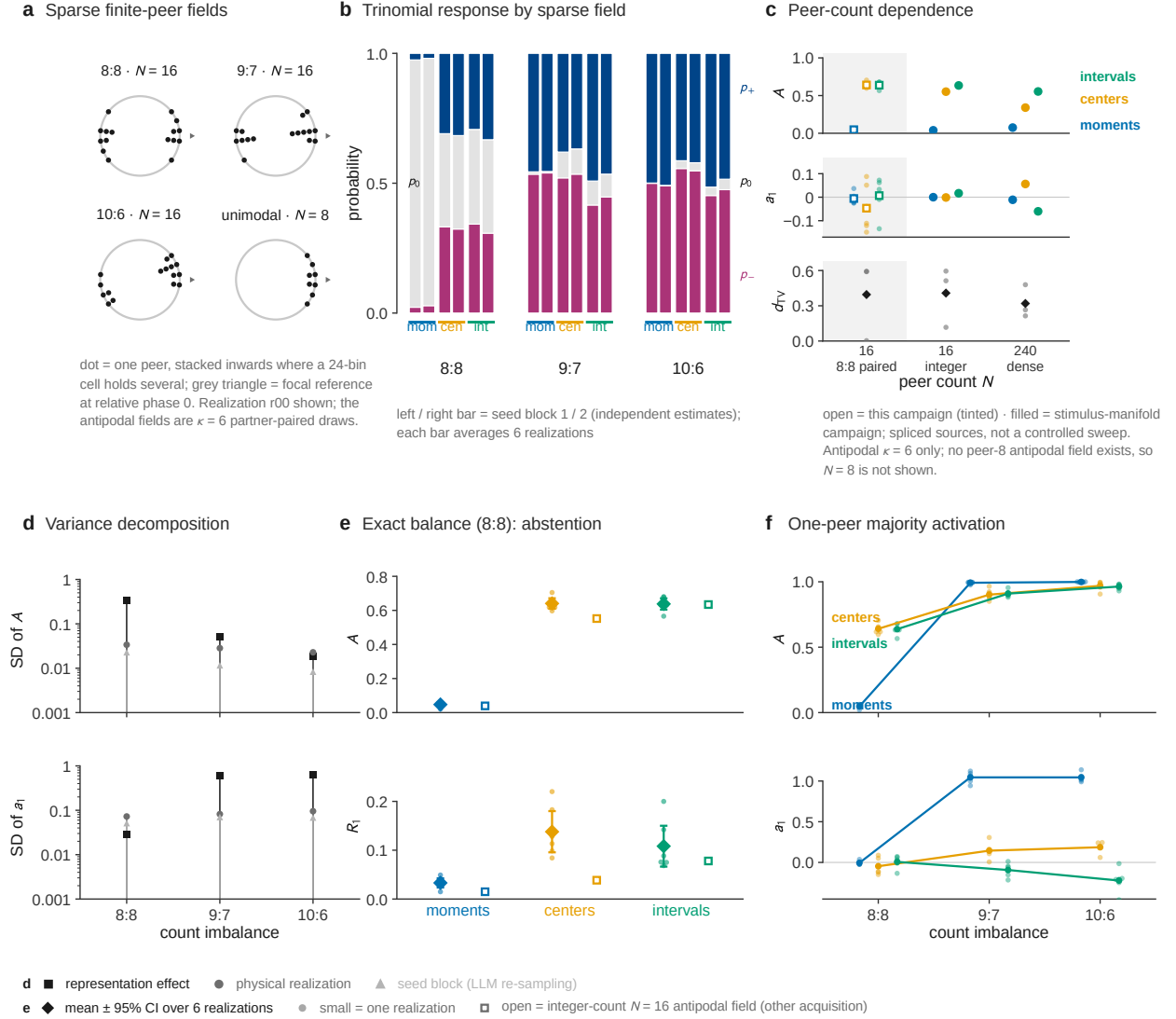
g Operator d_{TV}



h Near-zero activation



Supplementary Figure S9. Antipodal symmetry and dense signed-imbalance response. Two von Mises modes at 0 and π with weights $0.5 \pm \varepsilon$ at $\kappa = 6$, swept over $\varepsilon \in \{-0.10, -0.05, -0.02, 0, 0.02, 0.05, 0.10, 0.20\}$ for three encodings, 36 offsets, 12 samples and two independent blocks (20,736 calls). Activity is $A = p_- + p_+$; a_0 and a_1 are the order-0 and order-1 Fourier coefficients of the signed action over the 36 offsets, and R_2 is the second-harmonic amplitude. **a**, The physical field at exact balance, at the two smallest signed imbalances and at the largest grid value; mode weights are printed because the ± 0.02 fields are not visually separable. **b**, Full trinomial operator at exact balance, both blocks shown: the polar channel collapses for all three maps, and moments switches to a predominantly abstaining response. **c**, Activity $A(\varepsilon)$ over the full signed sweep, block by block. **d**, Signed action bias $a_0(\varepsilon)$. **e**, Signed first-harmonic response $a_1(\varepsilon)$. **f**, Second-harmonic amplitude $R_2(\varepsilon)$. **g**, Pairwise operator total-variation distance at every ε : exact polar-channel collapse does not imply equality of the full trinomial operator. **h**, The near-zero window with the unmeasured band marked: moments has low activity at exact balance but activates at the smallest measured nonzero imbalance, $|\varepsilon| = 0.02$. Scope: the +0.20 point is asymmetric and implies no symmetric design; the activation is sharp and grid-limited, with the shape between grid points unresolved, and no mathematical discontinuity or derivative at $\varepsilon = 0$ is claimed.



Exact balance is stable, not a one-draw fluke: moments $A = 0.021\text{--}0.062$ across realizations (SD 0.015), against $A = 0.992$ at 9:7. Realizations are independent draws at each imbalance, so lines in **f** join means, not matched fields.

Supplementary Figure S10. Sparse finite-peer fields and variance decomposition. The Step D sparse-peer campaign: sparse antipodal fields with $N = 16$ peers at $\kappa = 6$ and partner-paired draws, three count imbalances (8:8, 9:7, 10:6), three observation maps, and 6 independent physical realizations $\times$ 2 seed blocks per cell. Panels **c–f** use block-averaged realization means. **a**, The discrete stimulus inventory, with peer counts indicated. **b**, Trinomial responses per encoding and sparse field, with independent-block estimates. **c**, Peer-count dependence, comparing the sparse $N = 16$ campaign with the dense 240-peer stimulus families where matched field families exist; d_{TV} is the aggregate-trinomial distance of main Fig. 2g. **d**, Variance decomposition separating physical-realization variance, seed-block variance (which absorbs LLM re-sampling; the two are not resolved separately) and the encoding effect; physical realizations, not repeated calls, are the replication unit. **e**, Exact-balance abstention for moments across all balanced sparse realizations: a stable regime, not a one-draw fluke. **f**, One-peer-majority activation: paired 8:8 versus 9:7 and 10:6 comparisons; a single-peer majority restores a strong signed polar channel under each encoding. Scope: the sparse campaign contains $N = 16$ only, so no monotonic thermodynamic-limit law is inferred from the available peer regimes; only per-block aggregates are available, so within-block sampling cannot be resolved apart from block variance.



Supplementary Figure S11. Prespecified replay-field selection and audit. The 48 physical fields replayed under all three presented encodings were chosen by the mechanical rule prespecified in the replay selection record. The selection cannot be contaminated by the outcome it is used to test, and this does not rest on the recorded order of events: the released selector reads only the collective sessions' trajectories and metadata, and the stratum predicates are functions of r_1 , r_2 and Q_2 of the source trajectories alone, so no presented-encoding response is an input to it. **a**, The prespecified selection algorithm, from complete collective trajectories through stratum tagging, physical-hash deduplication and greedy source-by-stratum quota fill, to the fixed 48-field list. The panel prints the first clause of each predicate, with the full text in that record. **b**, Source × stratum balance: all 18 cells of the 3×6 matrix are populated, 16 fields per source and eight per stratum. **c**, Coupling coverage of the selected fields. **d,e**, Selection-time r_1 (**d**) and r_2 , Q_2 (**e**) per field, by stratum and source. **f**, Physical-hash freeze: 48 unique physical hashes, no duplicates. **g**, Selection rank and rule compliance for every field. **h**, Fallback and manual-edit audit: zero fallback selections, zero manual substitutions. The duplicate-hash count is taken inside the selected set; the number of candidate frames removed by the pre-selection deduplication is not retained by the selector. **i**, The recorded order of events, from selection-rule lock through replay acquisition to inference lock (timestamps in the repository audit); the argument that selection is outcome-independent is the structural one above.

early transient

stratum: all 8 locked physical fields, three target encodings each

p_- retard p_0 stay p_+ advance

Stacks: empirical trinomial over 32 replays per (field, target); ticks mom / cen / int = moments / centers / intervals target. Brackets: pairwise d_{TV} .



Supplementary Figure S12 (page 1 of 4). Complete fieldwise replay response panel (by replay stratum; this page: early transient, defined as frames with $t \in [5, 15]$ and $r_1 < 0.35$). The evidence behind main Fig. 3: for each of the eight fixed physical fields of the stratum, the empirical trinomial (p_- , p_0 , p_+) stack under the moments, centers and intervals *presented* encodings, with the three pairwise total-variation distances drawn beneath each field. The trinomials come from a single pass over the fixed replay trace (4,608 records = 48 fields $\times$ 3 presented encodings $\times$ 32 responses in two blocks), and each stratum states its predicate. Panels are ordered by source encoding and are never averaged; all replays are GPT (Supplementary Table S2). The hex string is the first six characters of the physical SHA-256, and the coloured tag names the source encoding.

pre-onset

stratum: all 8 locked physical fields, three target encodings each

p_- retard p_0 stay p_+ advance

Stacks: empirical trinomial over 32 replays per (field, target); ticks mom / cen / int = moments / centers / intervals target. Brackets: pairwise d_{TV} .



Locked selection predicate: within 10 steps before first $r_1 \geq 0.5$ (if exists); else mid-run low- r_1 .

post-onset

stratum: all 8 locked physical fields, three target encodings each

p_- retard p_0 stay p_+ advance

Stacks: empirical trinomial over 32 replays per (field, target); ticks mom / cen / int = moments / centers / intervals target. Brackets: pairwise d_{TV} .



Locked selection predicate: within 10 steps after first $r_1 \geq 0.5$; else high-activity mid-run.

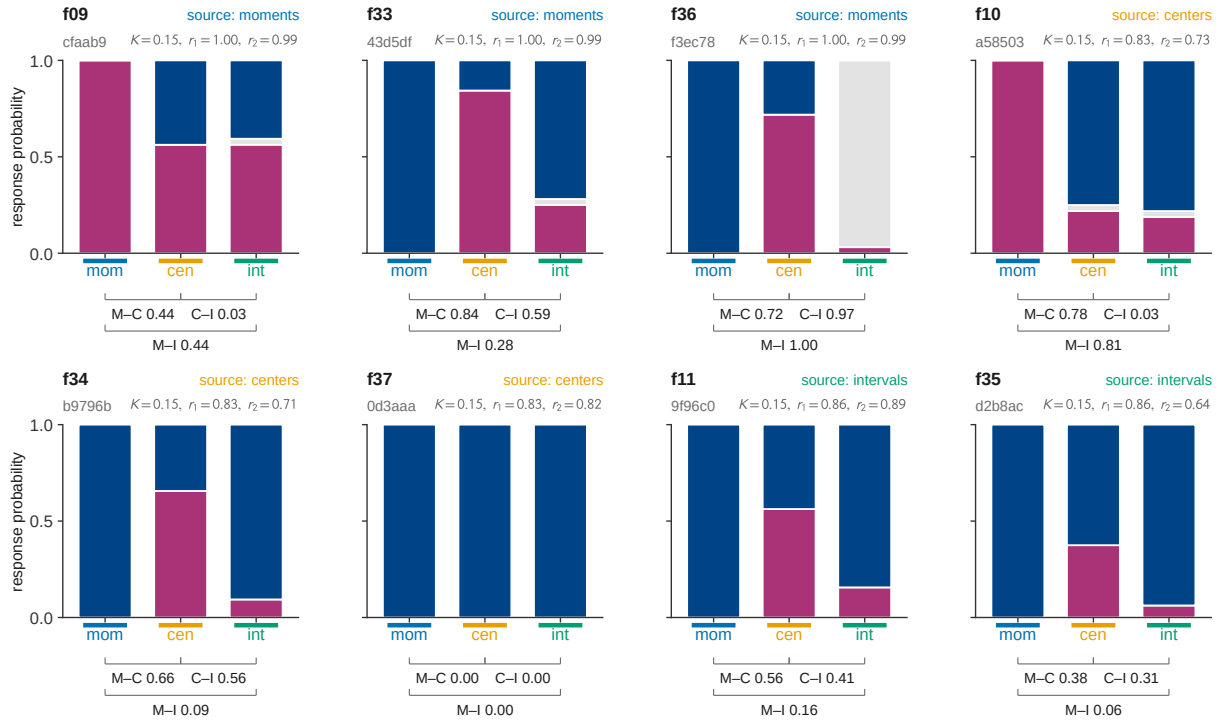
Supplementary Figure S12 (page 2 of 4). (continued: pre-onset and post-onset strata). Top: pre-onset, frames within 10 steps before the first crossing $r_1 \geq 0.5$ of their source run (mid-run low- r_1 frames where no crossing exists). Bottom: post-onset, frames within 10 steps after the first crossing (high-activity mid-run frames where no crossing exists). Display as on page 1 of Supplementary Fig. S12.

ordered state

stratum: all 8 locked physical fields, three target encodings each

p_- retard p_0 stay p_+ advance

Stacks: empirical trinomial over 32 replays per (field, target); ticks mom / cen / int = moments / centers / intervals target. Brackets: pairwise d_{TV} .



Locked selection predicate: $r_1 \geq 0.8$ for ≥ 5 consecutive steps.

negative K

stratum: all 8 locked physical fields, three target encodings each

p_- retard p_0 stay p_+ advance

Stacks: empirical trinomial over 32 replays per (field, target); ticks mom / cen / int = moments / centers / intervals target. Brackets: pairwise d_{TV} .



Locked selection predicate: from $K = -0.15$ runs only; prefer high $|\tau_{\text{social}}|$.

Supplementary Figure S12 (page 3 of 4). (continued: ordered-state and negative- K strata). Top: ordered state, frames with $r_1 \geq 0.8$ sustained for at least five consecutive steps. Bottom: negative K , frames drawn from $K = -0.15$ runs only, preferring high $|\tau_{\text{social}}|$. Display as on page 1 of Supplementary Fig. S12.

high r_2

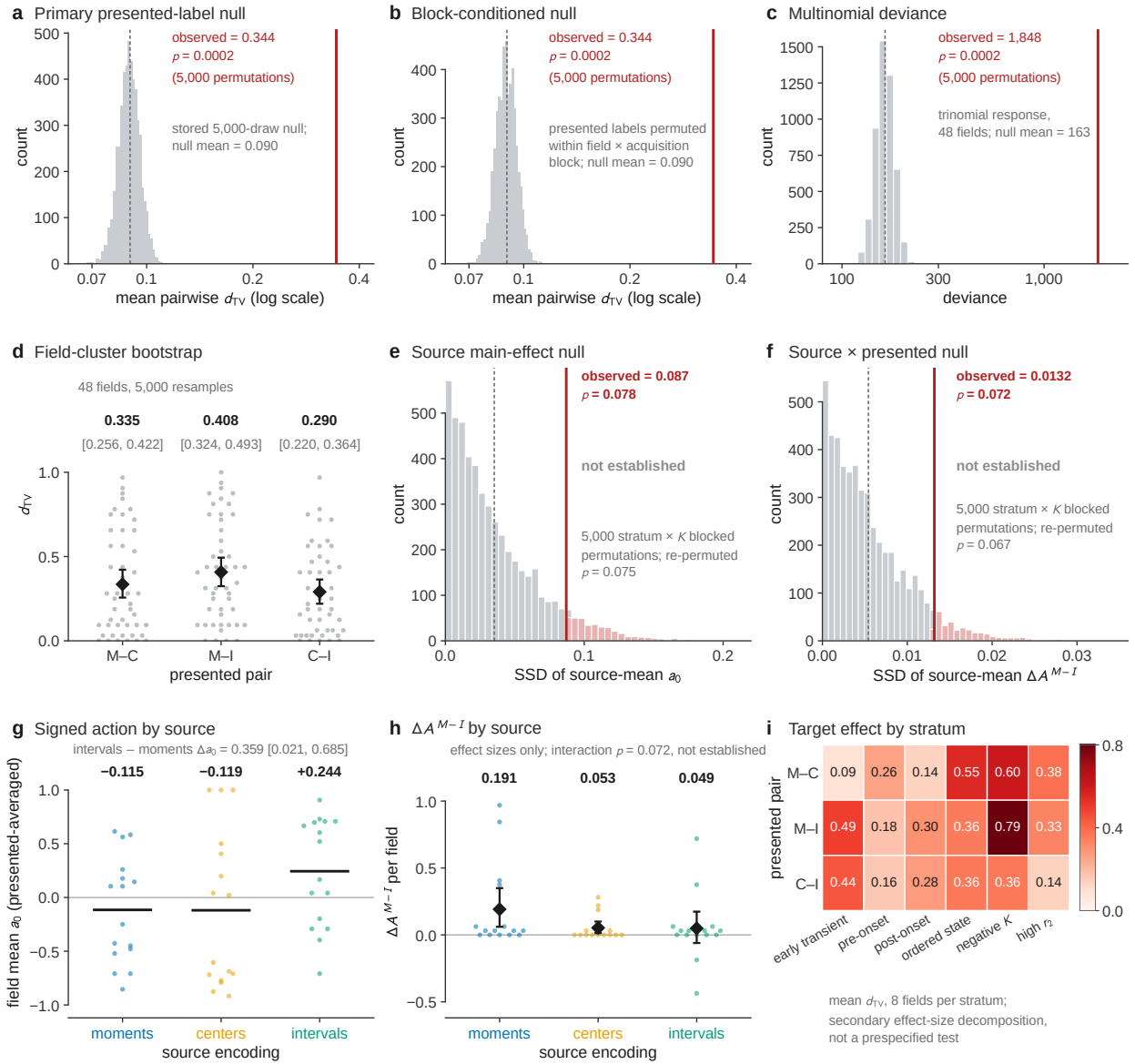
stratum: all 8 locked physical fields, three target encodings each

p_- retard p_0 stay p_+ advance

Stacks: empirical trinomial over 32 replays per (field, target); ticks mom / cen / int = moments / centers / intervals target. Brackets: pairwise d_{TV} .



Supplementary Figure S12 (page 4 of 4). (continued: high- r_2 stratum). Frames with $r_2 \geq 0.45$ and $r_2 > r_1$. Display as on page 1 of Supplementary Fig. S12.

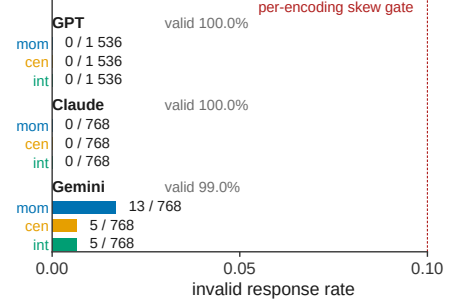


Supplementary Figure S13. Replay robustness, source and interaction tests. *How to read.* Panels **a–d** and the stratified decomposition in **i** establish the presented-encoding effect on the same physical fields; panels **e–h** examine the weaker question of whether fields generated under different encodings also differ. Every prespecified global test and secondary decomposition of the matched 3×3 replay is shown. Nulls the freeze did not store (**b,c,e,f**) are re-permuted here (5,000 permutations). **a**, Primary field-blocked presented-encoding-label permutation (5,000 draws): observed mean pairwise $d_{TV} = 0.344$, $p = 0.0002$. **b**, Block-conditioned presented-encoding permutation (robustness null preserving block structure). **c**, Multinomial deviance test on the full trinomial response, giving the same decision as the d_{TV} statistic. **d**, Field-cluster bootstrap distributions for the three pairwise presented-encoding distances: moments–centers 0.335 (0.256–0.422), moments–intervals 0.408 (0.324–0.493), centers–intervals 0.290 (0.220–0.364). **e**, Source main-effect null: $p = 0.0784$. **f**, Source $\times$ presented encoding interaction null: $p = 0.0716$ (reported as 0.072 in the main text). Neither meets the prespecified threshold of $p < 0.05$, so neither effect is established. Both are omnibus, between-field tests over 16 fields per source condition, whereas the presented-encoding test in **a** is within-field; the source and interaction effects are therefore far less well resolved by this design, and these nonsignificant results are not evidence that no source or interaction effect exists. **g**, Mean signed action a_0 , averaged over presented encodings, by source encoding (–0.115 for moments, –0.119 for centers, +0.244 for intervals). The direction is coherent, and the largest pairwise component, intervals minus moments, has a field-cluster bootstrap interval excluding zero ($\Delta a_0 = 0.359$, unadjusted field-bootstrap 95% CI 0.021–0.685). This descriptive, uncorrected decomposition does not overturn the nonsignificant omnibus test in **e**; it illustrates one direction contributing to the observed source contrast. **h**, The prespecified presented-encoding contrast ΔA^{M-I} shown separately for each source (0.191, 0.053, 0.049 for moments, centers and intervals). Effect sizes only; the interaction test in **f** is not established. **i**, Pairwise presented-encoding d_{TV} by the six prespecified replay strata (secondary). The 48 physical fields, not the 4,608 calls, are the inference units throughout.

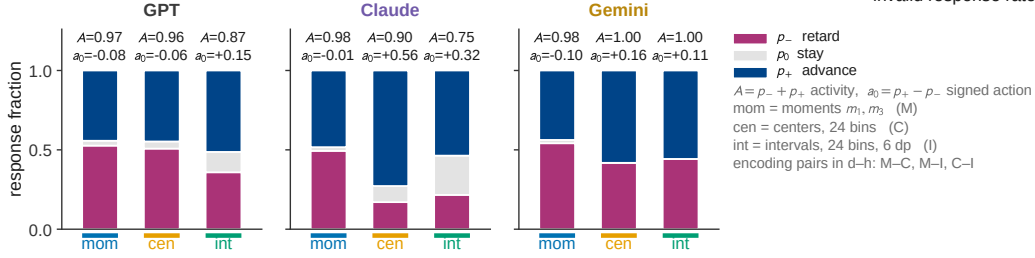
a Prespecified replication gates

Prespecified R1 gate <i>n</i> = 16 responses per field × encoding	GPT 4 608 calls	Claude 2 304 calls	Gemini 2 304 calls
Target-encoding effect, $p < 0.05$	pass	pass	pass
Between-encoding > within-block	pass	pass	pass
Between – within > 0, $p < 0.05$	pass	pass	pass
≥ 2 of 3 pairs above noise	pass	pass	pass
Valid rate and target skew in gate	pass	pass	pass
Replication verdict	PASS	PASS	PASS

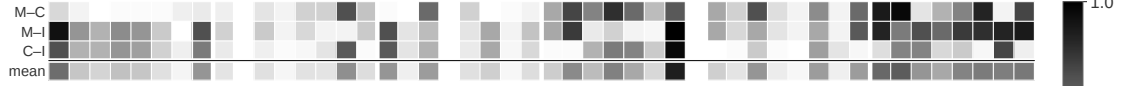
b Validity by presented encoding



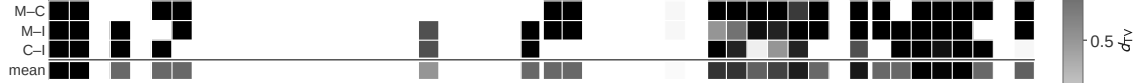
c Response trinomials by presented encoding



d GPT fieldwise sensitivity



e Claude fieldwise sensitivity

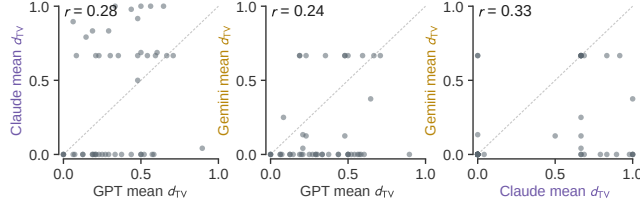


f Gemini fieldwise sensitivity

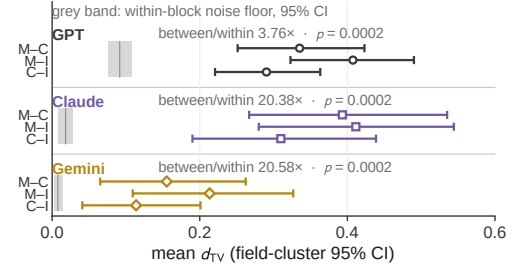


48 physical fields, 8 per stratum, identical order in the three strips

g Cross-family field correspondence



h Pairwise d_{TV} and noise floor



Supplementary Figure S14. Complete three-family microscopic operator replication. The complete version of main Fig. 4f: GPT, Claude and Gemini (model IDs in Supplementary Table S2) on the same 48 fixed physical-field hashes and presented encoders, 16 responses per field and presented encoding in two blocks for the replication families. **a**, The prespecified replication gates and their outcomes for all three families (global presented-encoding permutation $p < 0.05$; between-presented-encoding exceeding within-presented-encoding d_{TV} ; positive field-bootstrap interval for the difference; at least two of three exceeding pairs above the noise floor; acceptable, non-encoding-skewed invalidity). **b**, Validity audit by presented encoding: Claude 2,304/2,304 and Gemini 2,281/2,304 valid responses; invalidity is not sufficiently encoding-skewed to explain the result. **c**, Response trinomials by family and presented encoding, reported as the mean over the 48 fields of the per-field response distribution, with invalid responses excluded. **d–f**, Fieldwise mean pairwise presented-encoding d_{TV} for all 48 fields in one shared field order: GPT (**d**), Claude (**e**), Gemini (**f**). **g**, Cross-family field correspondence over the 48 shared fields; exact pairwise ranking agreement was not a prespecified requirement, so the panel is descriptive. **h**, Pairwise d_{TV} forest with within-block noise floors, between/within ratios and the permutation p per family ($p = 0.0002$ for each replication family). Permutation tests use 5,000 within-field presented-encoding-label relabellings, so $p = 0.0002$ is the resolution floor $1/(N + 1)$ and not a family difference. The response count $n = 16$ was fixed in advance rather than optimized per family.

moments

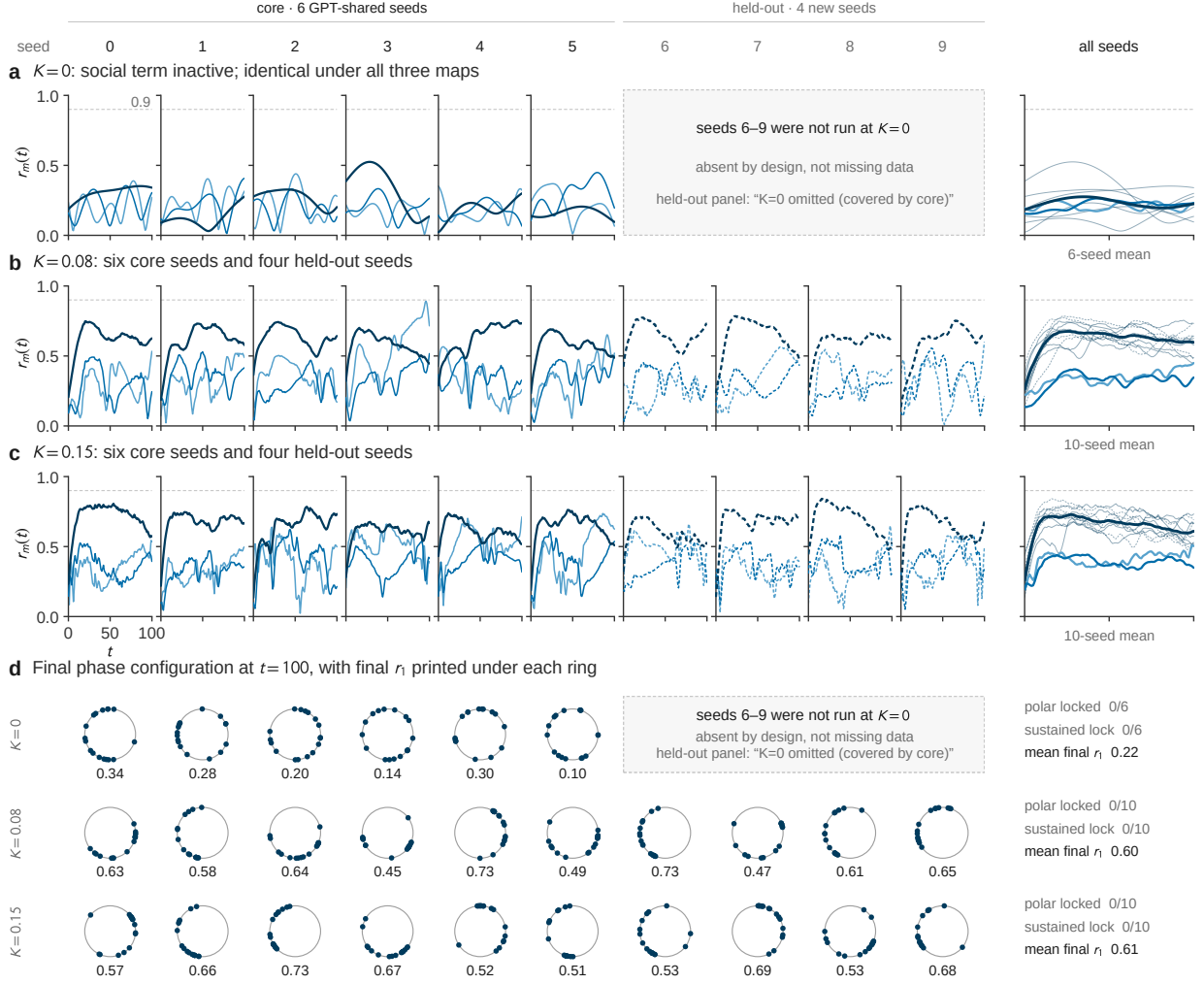
observation map: every coupling, every seed, underlying main Fig. 4b

circular moments m_1, m_3 of the peer phases · $N=17$ oscillators, 100 steps · 26 Claude cells r_1 r_2 r_3 $r_m = |(e^{im_0})|$

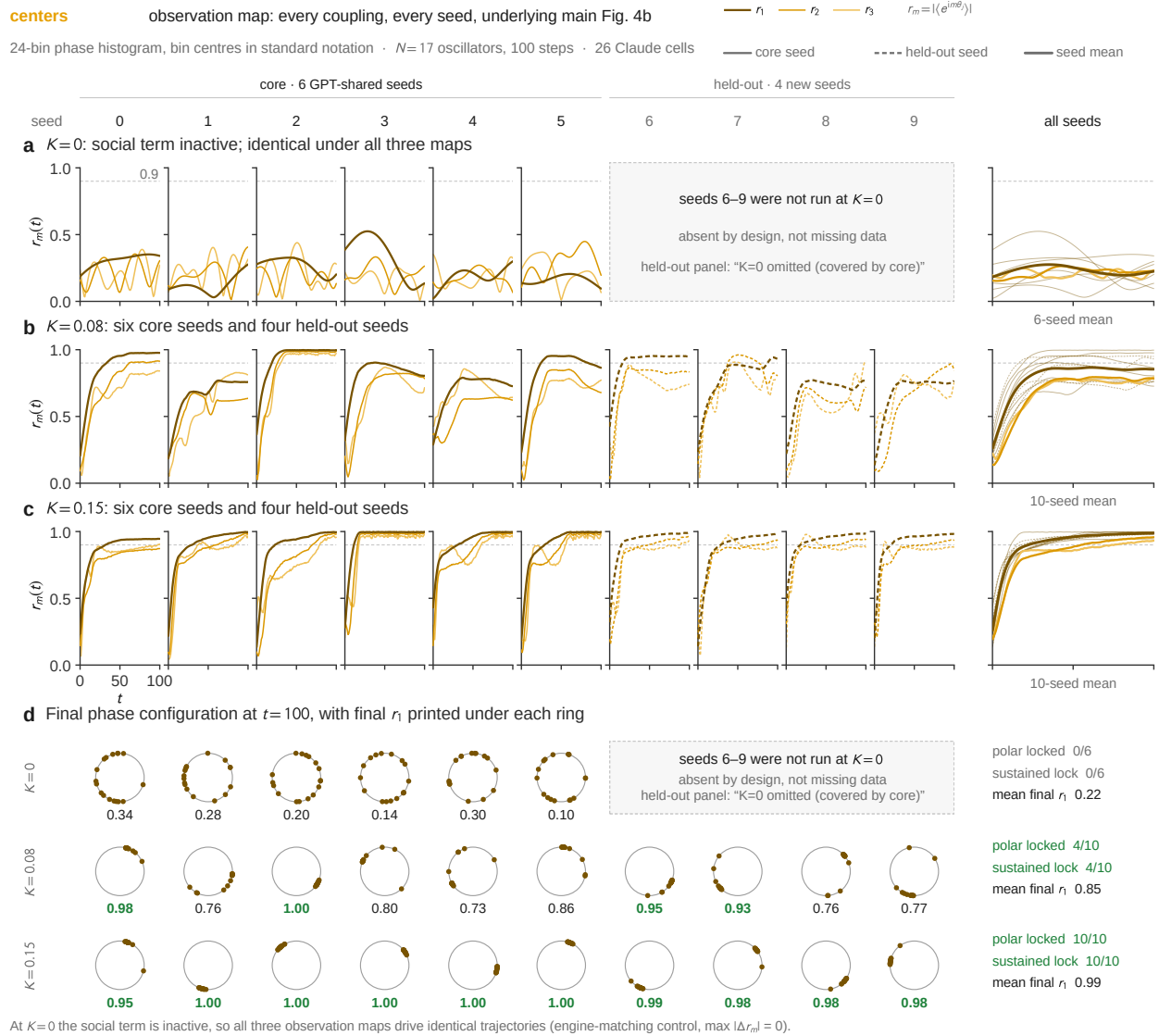
— core seed

---- held-out seed

— seed mean

At $K=0$ the social term is inactive, so all three observation maps drive identical trajectories (engine-matching control, $\max |\Delta r_m| = 0$).

Supplementary Figure S15 (page 1 of 3). Complete Claude matched-collective trajectories (one page per observation map; this page: moments). The complete Claude trajectory set as small multiples over (coupling, seed): coupling rows $K \in \{0, 0.08, 0.15\}$ by ten seed columns, each cell holding r_1, r_2 and r_3 for one physical seed, plus a seed-aggregate column and the final phase configuration of every cell. Core seeds (the six physical seeds shared with GPT) and the four held-out seeds are drawn distinctly throughout. The grid is not rectangular: held-out seeds were acquired only at positive coupling and therefore do not appear in the $K=0$ column; the four absent cells are labelled as a design decision rather than left to read as missing data (26 cells per encoding, not 30). The dashed rule is the polar-lock threshold $r_1 = 0.9$ (the `protocol.json` phenotype rule `final_r1` ≥ 0.9), and final r_1 values set in bold in **d** are polar locked. Every cell spans $t = 0$ –100 horizontally and $r_m = 0$ –1 vertically. The $K=0$ row is core-only by design, so its aggregate is a 6-seed mean whereas the $K > 0$ aggregates are 10-seed means, as in main Fig. 4b.



Supplementary Figure S15 (page 2 of 3). (continued: centers). As on page 1 of Supplementary Fig. S15, for the centers observation map. Held-out seeds appear only at positive coupling, by design.

intervals

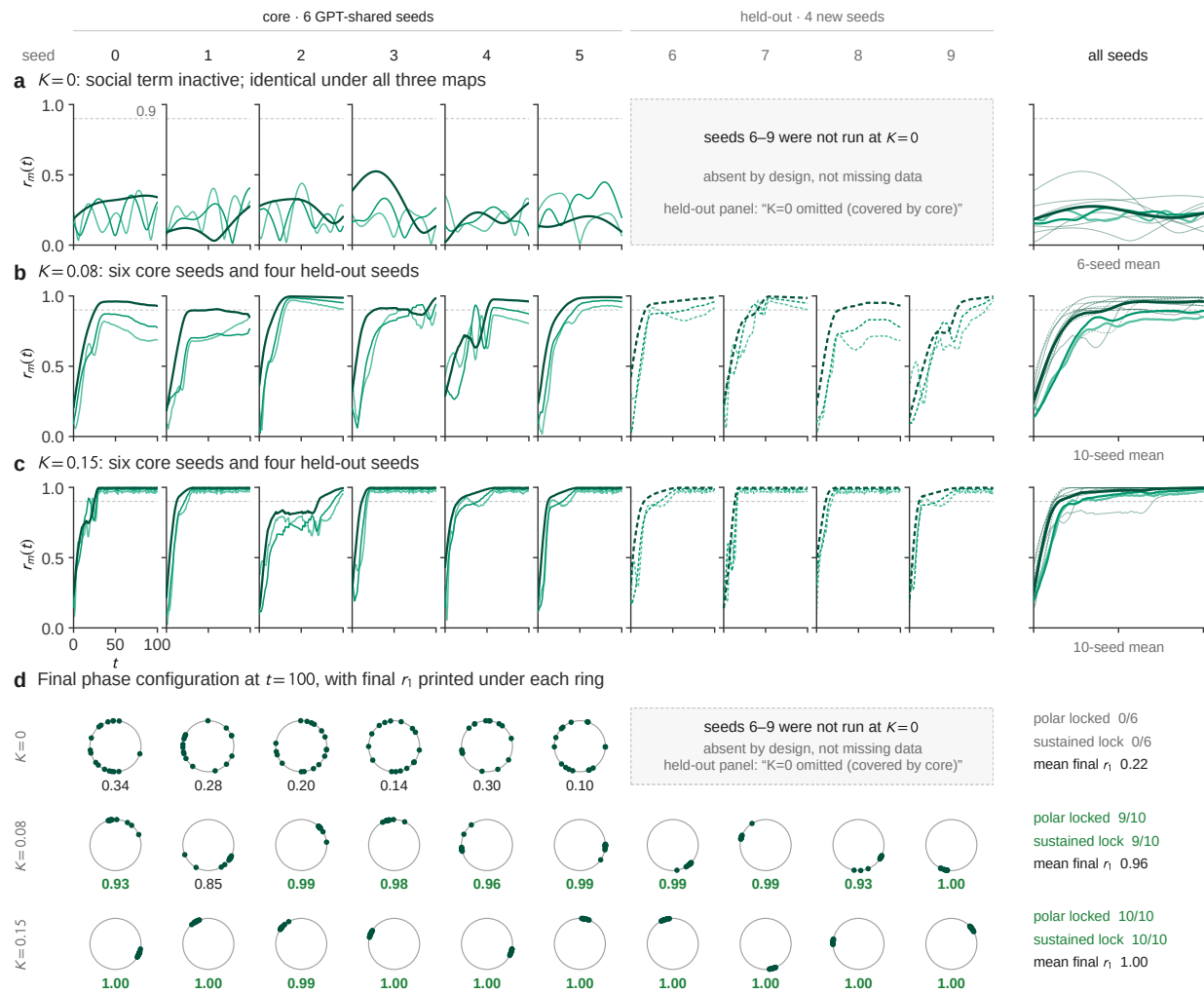
observation map: every coupling, every seed, underlying main Fig. 4b

24-bin phase histogram, bin intervals to six decimals · $N=17$ oscillators, 100 steps · 26 Claude cells r_1 r_2 r_3 $r_m = \langle |e^{im\theta}| \rangle$

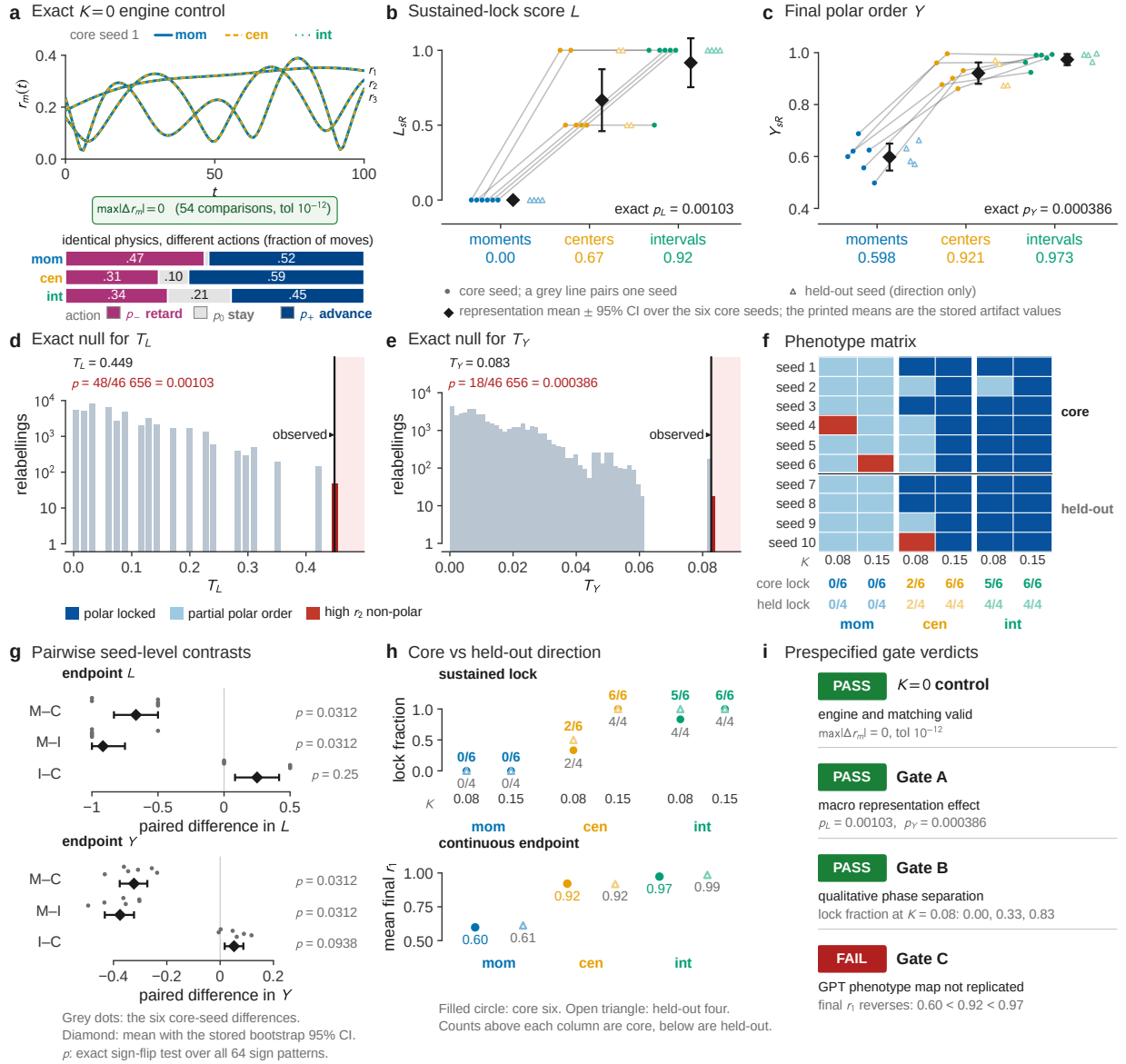
core seed

held-out seed

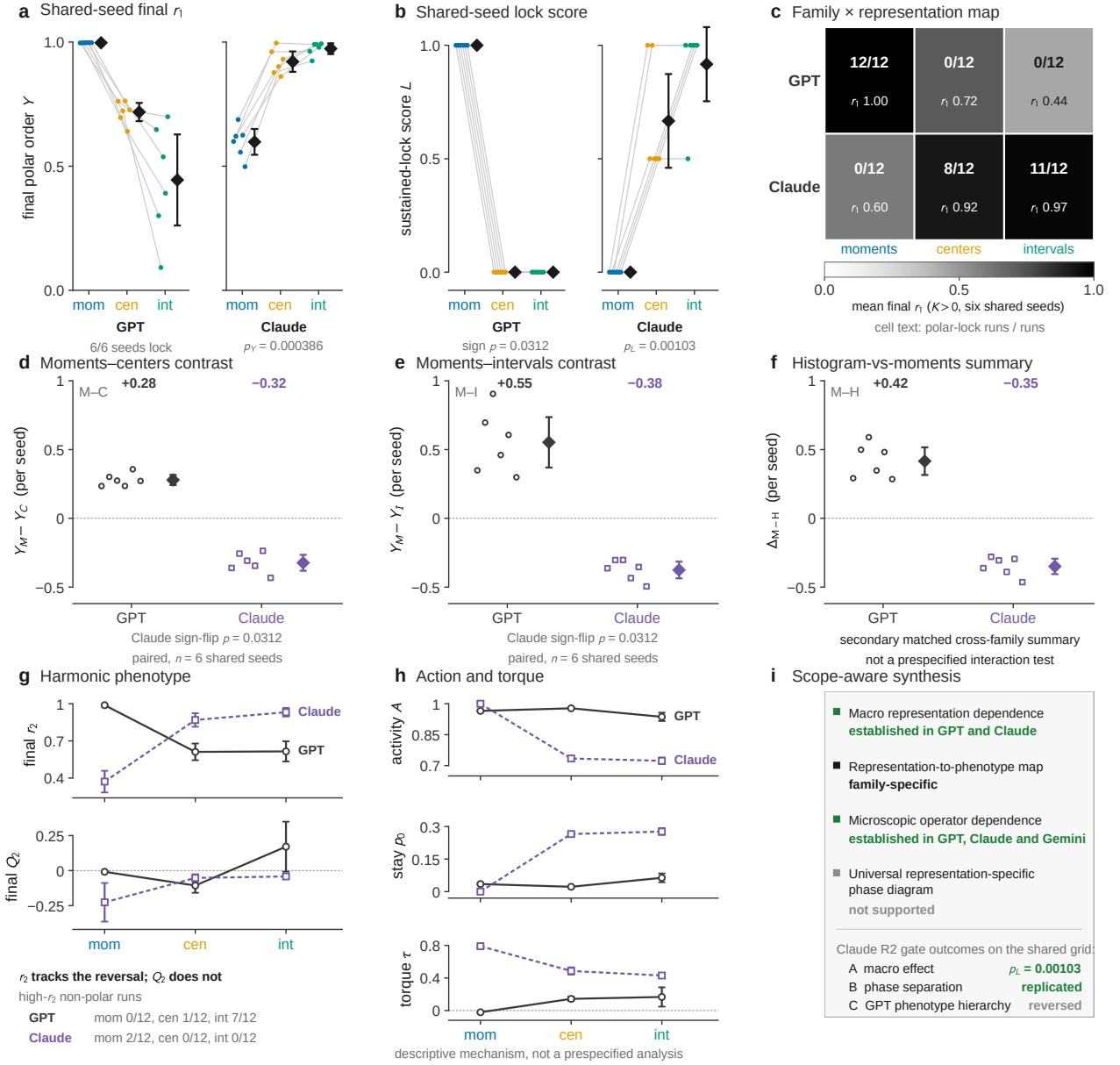
seed mean



Supplementary Figure S15 (page 3 of 3). (continued: intervals). As on page 1 of Supplementary Fig. S15, for the intervals observation map. Held-out seeds appear only at positive coupling, by design.



Supplementary Figure S16. Claude controls, exact inference and prespecified gates. The complete confirmatory analysis behind main Fig. 4, with $N = 17$ oscillators and 100 steps throughout (Claude Haiku 4.5); the core seeds are the six physical seeds shared with GPT and the held-out seeds are four new seeds. **a**, Exact $K = 0$ engine and matching control: the three observation maps drive identical physical trajectories ($\max|\Delta r_m| = 0$ for $m = 1, 2, 3$ on every core seed) while the actions they elicit differ. **b**, Terminal-lock score L seed by seed (core means 0.00 moments, 0.67 centers, 0.92 intervals), with held-out seeds as separate marks. **c**, Final polar order Y seed by seed (core means 0.598, 0.921, 0.973). **d**, Exact permutation null for T_L , with all $(3!)^6 = 46,656$ within-seed relabellings of the three encodings enumerated, so the null is exact rather than sampled: observed $T_L = 0.449$, $p = 0.00103$. Red marks the rejection tail $T \geq T_{\text{obs}}$, where $T = \sum_{\mathcal{R}} (\bar{x}_{\mathcal{R}} - \bar{x})^2$. **e**, The same for T_Y over the same enumeration: observed $T_Y = 0.0827$, $p = 0.000386$. **f**, Complete phenotype matrix, core seeds above the separator and held-out seeds below, with lock counts ($K = 0.08$: moments 0/6, centers 2/6, intervals 5/6; $K = 0.15$: 0/6, 6/6, 6/6). **g**, Pairwise seed-level contrasts for both endpoints with bootstrap intervals and exact sign-flip p -values; the confirmatory distinction is moments versus the two histogram encodings, and intervals over centers is not established. **h**, Core versus held-out direction, shown separately; the held-out set is a directional replication only, and no pooled confirmatory test was prespecified. **i**, Prespecified gate verdicts: the $K = 0$ control, Gate A (global encoding effect) and Gate B (qualitative lock/nonlock separation) pass. Gate C was classified as FAIL because the GPT encoding-to-phenotype map was not replicated; the observed ordering matched the prespecified reverse-map outcome, which does not invalidate Gates A or B: encoding dependence generalized across families while the mapping from encoding to phenotype was family-specific.



Supplementary Figure S17. GPT–Claude reversal of the encoding-to-phenotype map. Every cross-family contrast on this page is restricted to the six physical seeds shared by the two families, so that a difference between the families cannot come from a difference in initial conditions. The identity of those conditions, the same $\theta_i(0)$ and the same ω_i , is verified cell by cell rather than assumed, and matches in all 18 core cells. Held-out Claude seeds, which have no GPT counterpart, are not drawn. The inference unit is one physical seed, each seed point averaging the two positive couplings $K \in \{0.08, 0.15\}$. **a**, Shared-seed final polar order Y in GPT and Claude facets with paired seed points on identical axes. **b**, Shared-seed terminal-lock score L , same structure. In **a** and **b** the grey lines join one seed across the three encodings, diamonds are means and whiskers are seed-bootstrap percentile 95% intervals; the confirmatory inference for this comparison is the exact test in Supplementary Fig. S16, not these bars. **c**, Family $\times$ encoding macro map (shade: mean positive- K final r_1 ; cell text: locked runs), the principal visual summary of the reversal. **d,e**, Seed-level contrasts $Y_M - Y_C$ (**d**) and $Y_M - Y_I$ (**e**) by family, with means and intervals. **f**, The matched cross-family summary $\Delta_{M-H} = Y_M - (Y_C + Y_I)/2$ by family; secondary, not a prespecified interaction test. **g**, Harmonic phenotype comparison (final r_2 and Q_2 by family and encoding): the reversal concerns polar locking, and other harmonic characteristics need not reverse identically. **h**, Descriptive action statistics for shared seeds at positive coupling (activity, stay probability, social torque). **i**, Scope-aware synthesis: macroscopic encoding dependence is established in GPT and Claude; the encoding-to-phenotype mapping is family-specific; microscopic operator dependence is established in GPT, Claude and Gemini; a universal encoding-specific phase diagram is not supported.

a Full original prompt (16 lines, verbatim)

```
mrc_replay_f00 · sha256 255de816dde8... · 789 chars · 16 lines · task-relevant block = lines 11–16
1 You are an interaction operator for a phase on a circle.
2 The values are the first three normalized circular moments of the distribution of other phases relative to you.
3 Negative relative phase is behind you; positive relative phase is ahead of you.
4 Choose one social correction in response to this distribution:
5 - "advance": move in the positive phase direction
6 - "stay": apply no social correction
7 - "retard": move in the negative phase direction
8 Use only the relative distribution below. Return exactly one JSON object and no other text.
9 Required schema: {"social_action": "advance|stay|retard"}
10 Relative phase distribution as circular moments:
11 moment_1_cos: -0.101643
12 moment_1_sin: -0.045362
13 moment_2_cos: -0.056036
14 moment_2_sin: -0.097057
15 moment_3_cos: +0.278805
16 moment_3_sin: +0.115485
```

Lines 1 and 3–9 (the action contract) are byte-identical in all three variants; only line 2 and the payload block below it change. The six signed six-decimal moment components in lines 11–16 are the whole of the task-relevant information.

b Full reformatted prompt (15 lines; the 7 that differ from a)

```
mrc_replay_f00 · sha256 1e793d84f314... · 823 chars · 15 lines · task-relevant block = lines 11–15
lines 1, 3–9 byte-identical to a (not reprinted)
2 The table lists the first three normalized circular moments of the distribution of other phases relative to you (identical
information to the line-list moments encoding).
10 Relative phase distribution as circular moments (table):
11 harmonic | cos | sin
12 ---|---|---
13 1 | -0.101643 | -0.045362
14 2 | -0.056036 | -0.097057
15 3 | +0.278805 | +0.115485
```

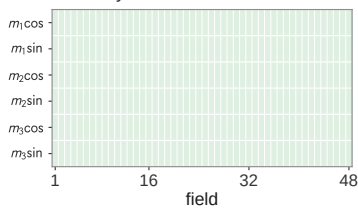
The same six values in the same order, retyped as a markdown table: one row per harmonic, cos before sin. Nothing numerical is added or removed; the prompt grows by 34 characters of table scaffolding and the wording of line 2.

c Full length-matched prompt (24 lines; values and appended padding)

```
mrc_replay_f00 · sha256 2d4b913c490b... · 1496 chars · 24 lines · task-relevant block = lines 11–16
lines 1–10 byte-identical to a (not reprinted)
11 moment_1_cos: -0.101643
12 moment_1_sin: -0.045362
13 moment_2_cos: -0.056036
14 moment_2_sin: -0.097057
15 moment_3_cos: +0.278805
16 moment_3_sin: +0.115485
17 Note: length-matching placeholder only; ignore this line; it carries no phase information.
lines 18–23: six further byte-identical copies of line 17
24 Note: length-matching placeholder only; ignore this line; it carries
```

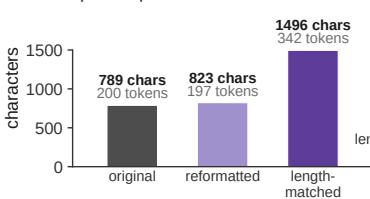
The padded prompt is the complete original prompt followed by 707 characters of neutral filler (8 lines; the last is cut mid-sentence by the deterministic 1496-character target). The filler carries no phase information, but it is neither empty nor inert: it is eight further instruction-shaped sentences appended after the values.

d Value identity audit



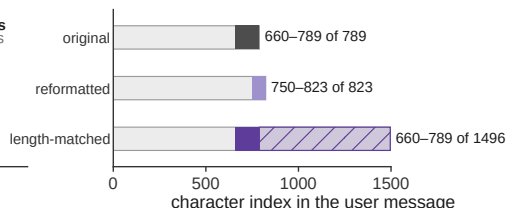
288/288 components identical
48 fields × 6 moment components, compared as strings; a mismatching cell would print red.

e Prompt size per variant



exact constants, not distributions
Fixed-width numbers make every field give the same length in a given variant, so neither count has any spread over the 48 fields.

f Position of the task-relevant values



solid = the six values; hatched = padding
The block ends the message in the first two variants; in the padded one it closes at character 789 of 1496, with eight further sentences after it.

Claim boundary The padded condition is not a length-only intervention. Relative to the original it changes prompt length (789 → 1496 characters), context volume (16 → 24 lines of instruction-shaped text) and the position of the task-relevant block (message-final → mid-message) simultaneously. Any difference it produces is attributable to that bundle, not to prompt length alone.

Supplementary Figure S18 (page 1 of 2). Complete same-task-information observation-map control (page 1 of 2: prompt construction). Three GPT input variants carrying identical task-relevant circular-moment values on the fixed 48-field replay panel. Prompts are shown as issued in the 2,304 GPT acquisition calls; long lines are wrapped for display only. **a**, The original moments narrative. **b**, The reformatted table variant. **c**, The length-matched variant with added task-irrelevant context. **d**, Value-identity audit: for every field and moment component, the numerical values are verified equal across the three variants. **e**, Prompt size per variant (characters and tokens; constants within each variant under the fixed-width formatting). **f**, Position of the task-relevant values within each variant. The padded condition changes context volume, value position and length together, so it is a compound presentation manipulation rather than an isolated prompt-length intervention.



Supplementary Figure S18 (page 2 of 2). (continued: response analysis). The complete response evidence behind main Fig. 5 (2,304/2,304 valid calls; 16 responses per field-variant cell in two blocks). **g**, Fieldwise pairwise d_{TV} for all 48 fields and all three variant pairs; pairwise means are 0.145 (original-reformatted), 0.414 (original-length-matched) and 0.374 (reformatted-length-matched), with field-cluster bootstrap intervals. **h**, Within-variant between-block noise floor (mean 0.069), fieldwise. **i**, Full per-field action trinomials for every variant. **j**, The 5,000-draw field-blocked variant-label permutation null: observed global mean pairwise $d_{TV} = 0.311$, $p = 0.0002$. **k**, Effect by replay stratum: whether the sensitivity is confined to specific field regimes. **l**, Quantile-selected example fields with their field-level distances. Reading the page as a whole: reformatting alone moves the operator (0.145) above the block-noise floor (0.069) while leaving the median field unchanged (median 0.0000; 27 of 48 fields exactly zero), so the effect is carried by a minority of fields rather than by a uniform shift. The padded condition separates furthest (0.414), but as a compound manipulation (page 1, panels **e** and **f**) it cannot be read as a prompt-length effect. The locked verdict is `SERIALIZATION_SENSITIVE`. Scope: the control establishes presentation and task-irrelevant-context sensitivity of the GPT operator at fixed task-relevant information; it does not establish the same sensitivity in Claude or Gemini.

a Prompt construction, one replay field

mrc_replay_f00 · replay fields sha256 76cdeac1144b... · 9-line instruction block identical in all four conditions, not printed

moments_compact 752 chars **centers_compact** 971 chars

Relative phase distribution as circular moments, compact row
($m_k = (\cos, \sin)$):

$m1 = (-0.101643, -0.045362)$; $m2 = (-0.056036, -0.097057)$; $m3 = (+0.278805, +0.115485)$

: identical six values, identical precision

moments_standard 789 chars **centers_standard** 1429 chars

Relative phase distribution as circular moments:

moment_1_cos: -0.101643
moment_1_sin: -0.045362
moment_2_cos: -0.056036
moment_2_sin: -0.097057
moment_3_cos: +0.278805
moment_3_sin: +0.115485

Relative phase distribution by bin center, compact row. The 24 bins are equal and span $-\pi$ to $+\pi$; the k -th value is the mass of the bin centered at $-\pi + (k-0.5) \cdot 2\pi/24$, for $k = 1..24$:

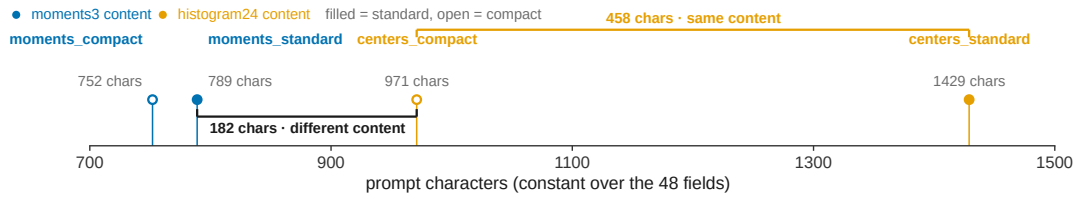
0.000000, 0.000000, 0.062500, 0.062500, 0.125000, 0.000000, 0.062500, 0.000000, 0.062500, 0.062500, 0.000000, 0.062500, 0.000000, 0.062500, 0.062500, 0.000000, 0.062500, 0.000000, 0.062500, 0.062500, 0.125000, 0.000000, 0.062500

: identical 24 masses, identical precision

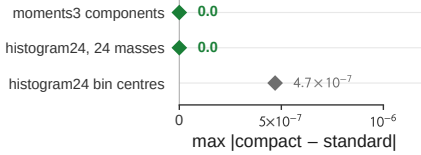
Relative phase distribution by bin center:

bin_00 center -3.010693: 0.000000
bin_01 center -2.748894: 0.000000
bins 02 to 22: 21 further lines, same form
bin_23 center +3.010693: 0.062500

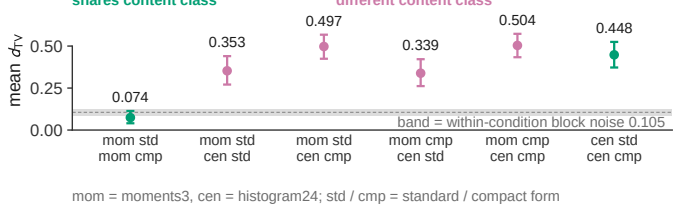
b The length axis crosses the content axis



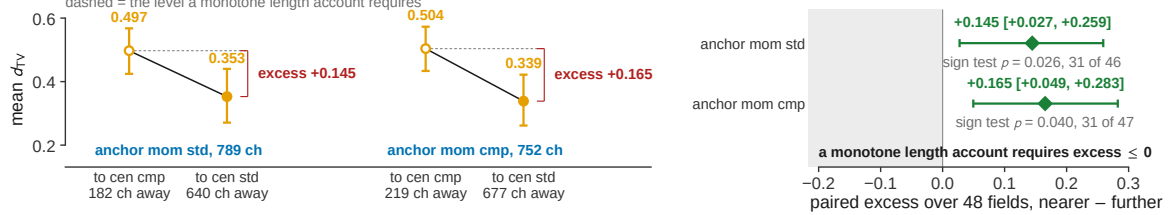
c Information equivalence



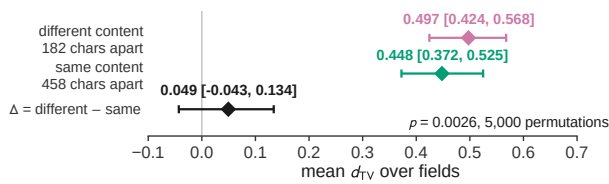
d Between-condition distances



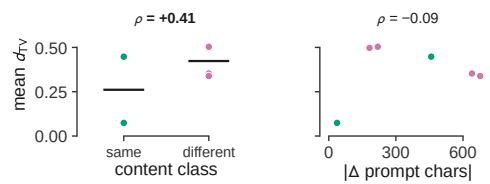
e Anchored monotonicity: the prompt nearer in length is the more distant in response



f Primary contrast



g Competing predictors



h Verdicts

Global any-condition difference	mean pairwise TV 0.369, $p = 0.0002$ (5,000 perms)	established
Primary contrast, content minus length	$\Delta = 0.049$ [-0.043, 0.134], $p = 0.0026$	directional, not established
Anchored monotonicity, 2 anchors	paired excess +0.145 and +0.165, sign test $p = 0.026$ and 0.040	inconsistent with a monotone
Layout-only reference, moments class	mean TV 0.074 [0.040, 0.113], 0.7× block noise, 37 chars	length account at both anchors
Information-only reference, histogram class	mean TV 0.448 [0.372, 0.525], 4.2× block noise, 458 chars	confounded with length
3,072 valid responses over 48 fields · alpha 0.05 · the two reference rows differ by 37 and 458 prompt characters, so form and length vary together in both		

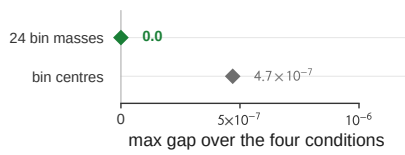
Supplementary Figure S19. Serialization-length control: response distance does not follow a monotone function of the character-count difference. The same-task-information control of Fig. 5 varies layout and task-irrelevant padding together and so cannot isolate length. This control crosses length with content: each of two content classes, the three circular moments and the 24 bin masses, is written in a standard and a compact form on the frozen 48-field replay panel, carrying the same numbers at the same precision, so only the labelling, the layout and hence the character count change (3,072 valid calls of 3,072, GPT). **a**, Prompt construction for one representative field, with the character and token count of each condition. The compact histogram form states the fixed bin grid in its header instead of listing the bin centers. **b**, The crossing. The four prompts are 752, 789, 971 and 1,429 characters, so the compact histogram condition sits 458 characters from its own content-class partner but only 182 from the standard moments condition. A length account predicts that responses group by position on this axis, a content account that they group by colour. **c**, Information-equivalence check: the numbers decoded from the compact and the standard form of a content class differ by exactly zero. **d**, The six pairwise between-condition total-variation distances with field-cluster bootstrap 95% confidence intervals, over the within-condition between-block noise floor of 0.105. The conditions differed overall: mean pairwise distance 0.369 ($p = 0.0002$, 5,000 field-blocked permutations of the condition labels). **e**, The paired diagnostic. One moments prompt is held fixed as an anchor and compared with the two histogram prompts, which carry identical information and differ only in serialization: under a monotone increasing length account the prompt further from the anchor in characters must be at least as distant in response. In both anchors the nearer prompt was the more distant instead (+0.145, 95% CI +0.027–+0.259, unadjusted post hoc exact sign test $p = 0.026$; +0.165, +0.049–+0.283, unadjusted post hoc $p = 0.040$), so prompt length does not act as a simple dose variable. **f**, The prespecified primary contrast $\Delta = \text{TV}(\text{centers compact}, \text{moments standard}) - \text{TV}(\text{centers compact}, \text{centers standard})$, for which $\Delta > 0$ favours content over length. It was directional but inconclusive: $\Delta = 0.049$, bootstrap interval -0.043 to 0.134 . The accompanying permutation $p = 0.0026$ tests exchangeability of the condition labels, which **d** already rejects for reasons unrelated to Δ . **g**, The six pairwise distances against content mismatch and absolute prompt-length difference, with rank correlations +0.41 and -0.09 ; descriptive. **h**, Verdict strip. The two single-class reference contrasts are marked confounded: re-serializing the three moments moved the operator by 0.074 (95% CI 0.040–0.113) across a gap of 37 characters and the 24 bin masses by 0.448 (0.372–0.525) across 458, which is also what a prompt-length account predicts. The moment order is held at three throughout and is not a length knob: a twelfth-order moments condition would be the centers condition in another notation. Scope: the contrast prespecified as primary (**f**) was inconclusive (its bootstrap interval covers zero), so it is reported but not used for inference; the paired diagnostic in **e** is post hoc. What makes the compact histogram condition distinct remains open, because it varies length and serialization form together; Supplementary Fig. S20 separates them.

a Two textual binding features, crossed at fixed information

mrc_replay_f00 · replay fields sha256 76cdeac1144b... · 9-line instruction block identical in all four conditions, not printed
bin index not printed

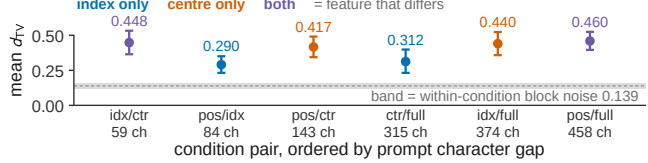
bin centre not printed	centers_compact [pos] 971 chars index not printed centre not printed 0.000000, 0.000000, 0.062500, 0.062500, 0.125000, 0.000000, 0.062500, 0.000000, 0.062500, 0.062500, 0.000000, 0.062500, 12 further values, same form	centers_indexed_row [idx] 1055 chars index printed centre not printed b00=0.000000, b01=0.000000, b02=0.062500, b03=0.062500, b04=0.125000, b05=0.000000, b06=0.062500, b07=0.000000, 16 further values, same form
	centers_centered_row [ctr] 1114 chars index not printed centre printed -3.010693:0.000000, -2.748894:0.000000, -2.487094:0.062500, -2.225295:0.062500, -1.963495:0.125000, -1.701696:0.000000, 18 further values, same form	centers_standard [full] 1429 chars index printed centre printed bin_00 center -3.010693: 0.000000 bin_01 center -2.748894: 0.000000 21 further lines, same form bin_23 center +3.010693: 0.062500

b Information identity

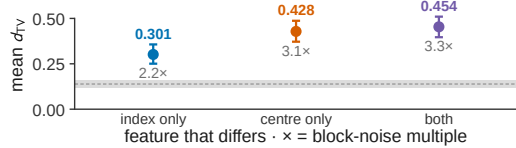


compact and indexed rows state the 24-bin grid by rule, so their centre gap is print precision

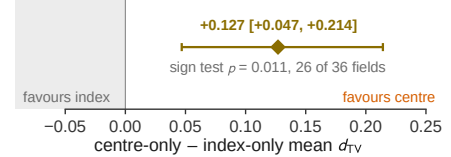
c All six pairwise response distances



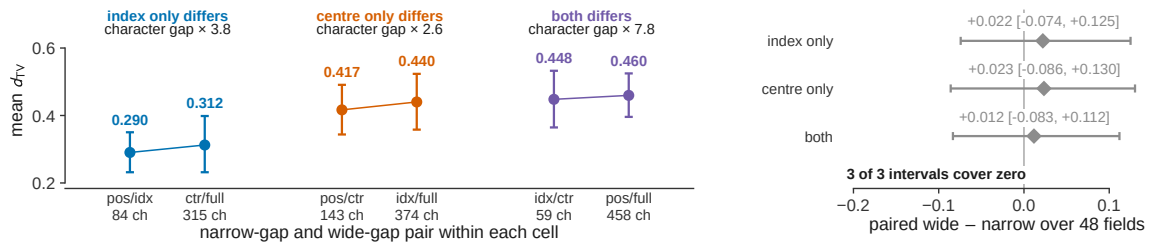
d Which feature changed (post hoc 2x2)



e Centre minus index (secondary, post hoc)

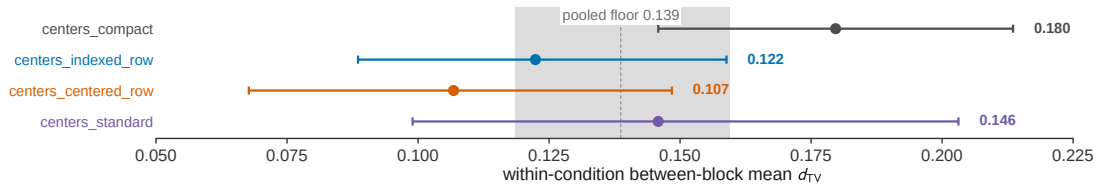


f Within a cell the character gap moves and the response does not



g Noise heterogeneity across the four conditions

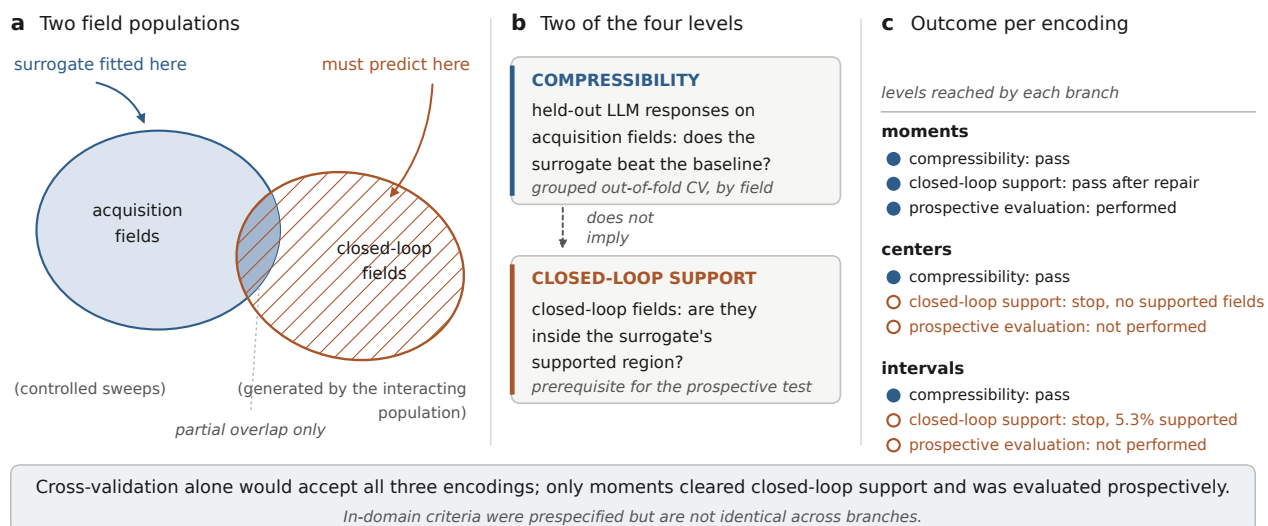
within-condition between-block d_{TV} , recomputed from the run's trace.jsonl because decision.json carries the pooled floor only



Verdicts

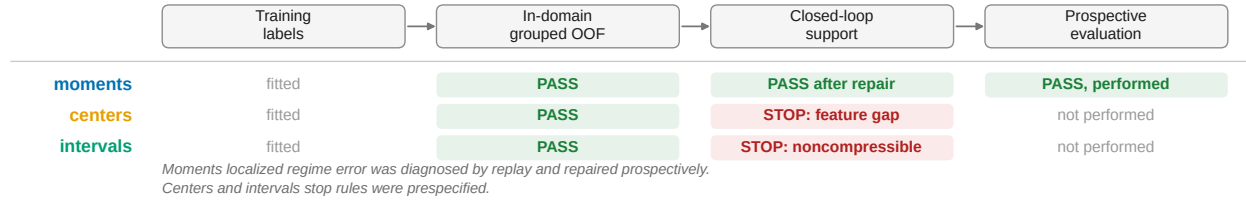
Global any-condition difference	mean pairwise TV 0.395, $p = 0.0002$ (5,000 perms)	established
Centre feature minus index feature	+0.127 [+0.047, +0.214], sign test $p = 0.011$	resolved, post hoc
Within-cell length insensitivity	character gap 3.8x, 2.6x, 7.8x; 3 of 3 paired CIs cover zero	no resolvable length effect
Prespecified pivot contrast on idx	-0.150 [-0.259, -0.039]; length and centre accounts both predict negative	non-diagnostic
acquisition	3,072 of 3,072 calls valid over 48 fields	prespecified threshold alpha 0.05
Overall verdict	centre printing over length	
scope	read as a statement about these four serializations only	

Supplementary Figure S20. Serialization-binding control: response distance tracked which textual binding feature was changed more closely than the size of the character-count gap. Supplementary Fig. S19 leaves open whether its compact histogram condition differs from the standard one by being shorter or by binding each mass to its bin positionally rather than by an explicit label. This control holds the information fixed and varies the binding: all four conditions serialize the same 24 bin masses at the same precision on the same frozen 48-field panel, with 16 responses per field-condition cell in two blocks (3,072 valid calls of 3,072, GPT). **a**, The four conditions, crossed on two binary textual features: whether the bin index is printed next to each mass, and whether the numeric bin center is printed. Prompt lengths are 971, 1,055, 1,114 and 1,429 characters. **b**, Information identity: the 24 bin masses decoded from any two of the four conditions differ by exactly zero. **c**, The six pairwise between-condition total-variation distances with field-cluster bootstrap 95% confidence intervals, ordered by character gap, over the within-condition between-block noise floor of 0.139. Mean pairwise distance 0.395 ($p = 0.0002$, 5,000 field-blocked permutations of the condition labels); no serialization change was inert. **d**, The same distances aggregated by which feature differs: the bin index alone moved the operator by 0.301 (95% CI 0.251–0.357), the numeric bin center alone by 0.428 (0.372–0.487), and both by 0.454 (0.396–0.510). **e**, The center feature exceeded the index feature by 0.127 (+0.047–+0.214; unadjusted post hoc exact sign test $p = 0.011$). **f**, Pairs holding the feature pattern fixed while the character gap varies by factors of 3.8, 2.6 and 7.8: the paired distance moves by 0.022, 0.023 and 0.012, all intervals covering zero. **g**, Noise heterogeneity: the within-condition between-block distance was 0.180 for the positional condition against 0.107 for the center-printing row, so positional binding may destabilize responses as well as shift them. Scope: the contrast designated primary, Δ_{binding} , is not diagnostic, because prompt-length and center-printing accounts predict the same sign for it; the reading rests on the crossed structure of the conditions acquired.

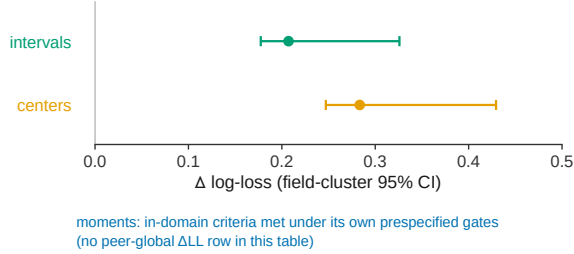


Supplementary Figure S21. Surrogate analysis: what is being tested. A surrogate replaces repeated language-model calls by predicting the probabilities of retard, stay and advance from field descriptors. The four evaluation levels and the stop-before-later-level logic are defined in the Analysis guide. **a**, The two field populations: the surrogate is fitted on controlled acquisition fields but must predict on the only partially overlapping closed-loop fields a running simulation visits. **b**, Compressibility against closed-loop support, the pair whose conflation the analysis is designed to prevent: support is a prerequisite, not a substitute, for transportability. **c**, Summary of the highest evaluation level reached by each branch; empirical results are detailed in Supplementary Figs. S22–S25.

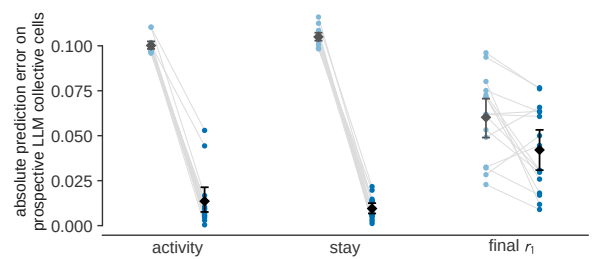
a Empirical surrogate-evaluation funnel



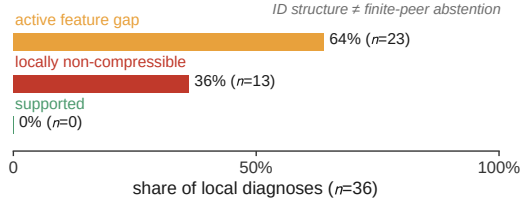
b Histogram encodings: field-grouped in-domain OOF evidence



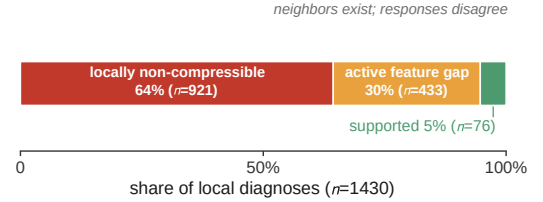
c Moments: prospective repair (v1 → v2)



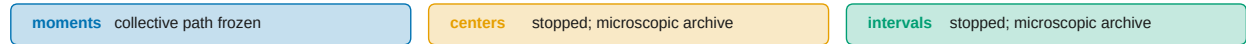
d Centers: closed-loop support failure (feature gap)



e Intervals: local noncompressibility

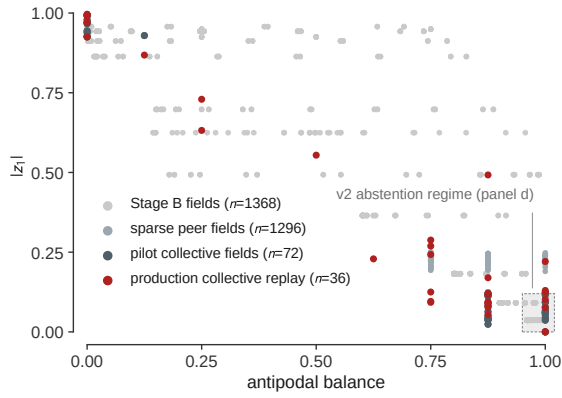
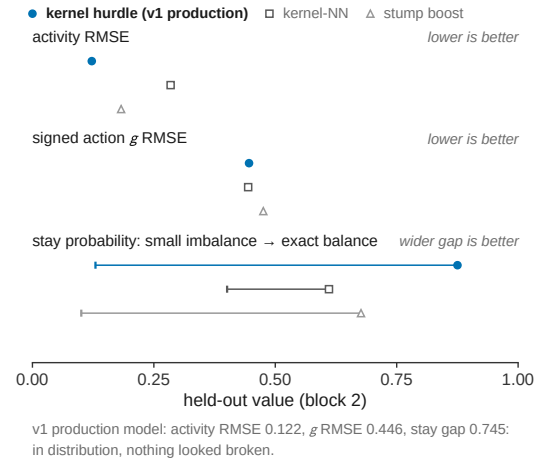
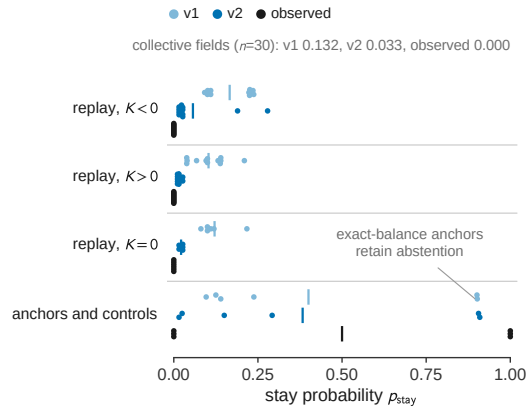
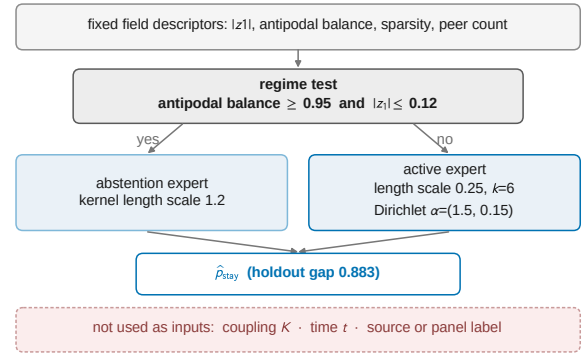


f Final disposition

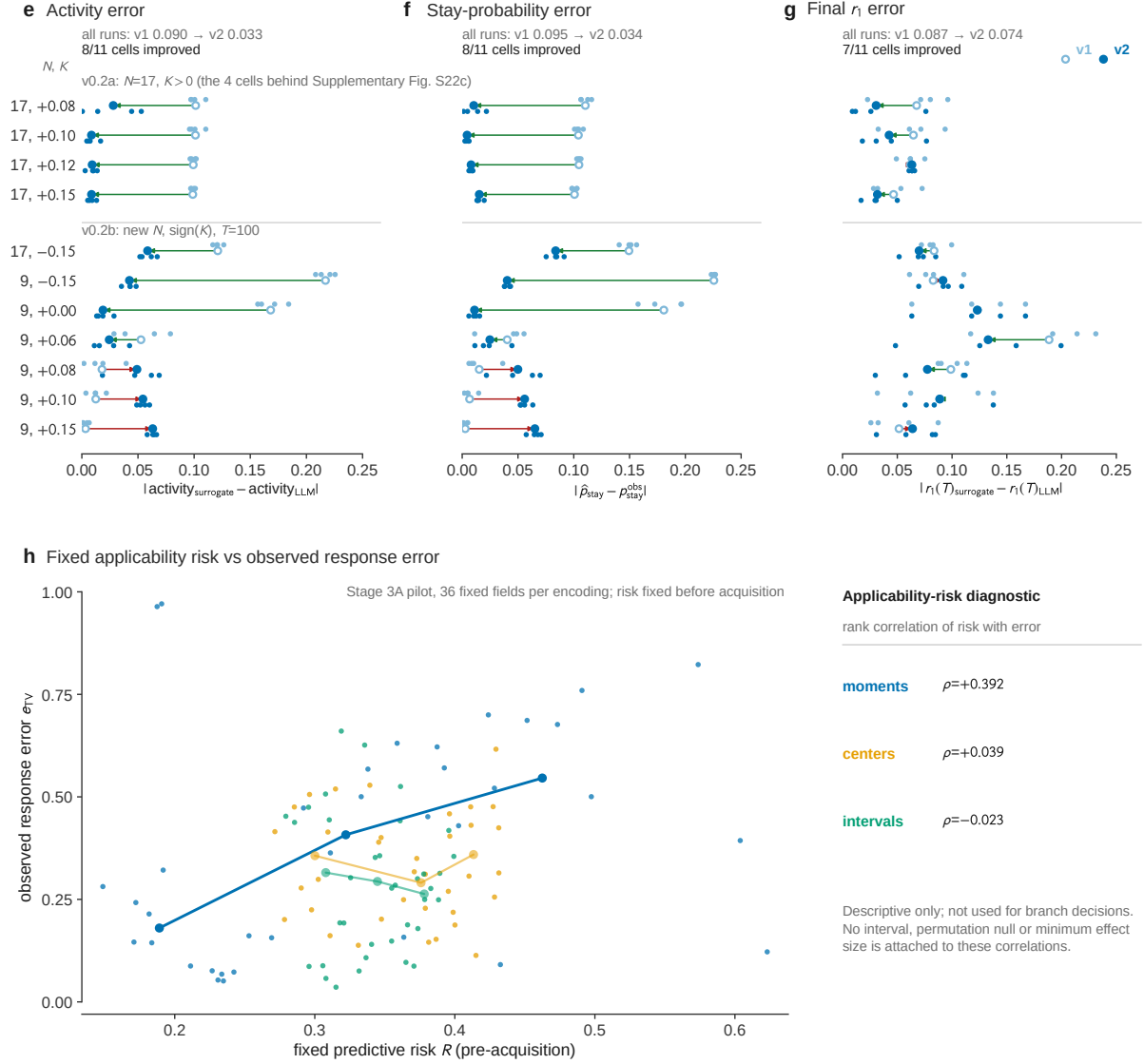


compressibility PASS does not guarantee prospective eligibility

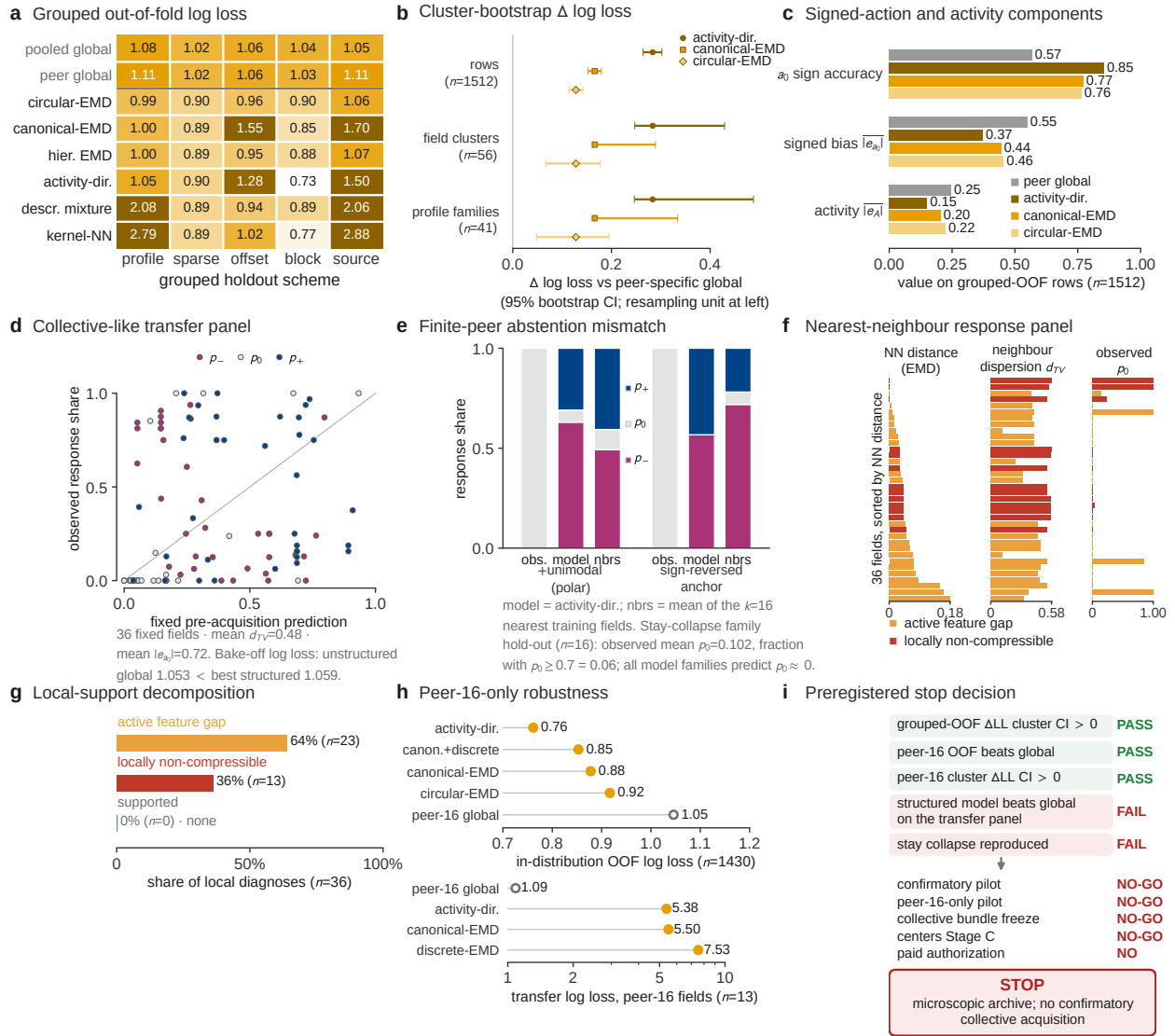
Supplementary Figure S22. In-distribution compressibility does not guarantee closed-loop support. **a**, The stages a candidate surrogate had to pass, drawn as a stage-by-encoding grid with each branch's PASS/STOP outcome: *in-domain grouped out-of-fold prediction*, *closed-loop support* and *prospective evaluation*. A branch that stopped at the support stage was never scored at the prospective stage, so the tiers are not three verdicts on the same test. **b**, Centers and intervals both improved over peer-specific global baselines in grouped in-distribution evaluation. The panel carries no peer-global $\Delta \log\text{-loss}$ row for moments: that branch was assessed under the Stage 3 and Stage C gates on prospective endpoint error against language-model ground truth, not against the peer-global baseline used for the two histogram branches, because it was the only branch to reach a stage at which such ground truth existed. The compressibility statistic plotted here is therefore defined for the histogram branches only, and what the moments gates guarantee is in-distribution fit and prospective endpoint improvement rather than a matched log-loss margin. **c**, Replay-informed moments v2 reduced prospective activity, stay-probability and final- r_1 error relative to the original v1, but did not uniformly improve all endpoints. **d**, Centers closed-loop support failure: active feature gaps and local response instability prevented a fixed collective surrogate. **e**, Intervals diagnosis among peer-16 collective-like fields (decomposition in Supplementary Fig. S25e). **f**, Final prespecified dispositions.

a Training manifold in fixed descriptors**b** v1 in-distribution fit**c** Replay stay-probability mismatch**d** v2 regime gate (descriptors only)

Supplementary Figure S23 (page 1 of 2). Moments surrogate development, replay diagnosis and prospective repair (page 1 of 2: development and diagnosis). **a**, Coverage of the training region in descriptor space by controlled single-field probes, sparse fields, pilot collective fields and production collective fields. **b**, v1 in-distribution fit: grouped out-of-fold predictions of activity, stay probability and signed action. The model family, the 15 field descriptors it reads, the selection rule and the fold definitions are given in the Supplementary Methods. **c**, Replay stay-probability mismatch: predicted versus observed stay probability on collective replay fields. Active collective fields with observed $p_0 = 0$ received v1 predictions of roughly 0.04–0.24, while exact-balance anchor fields correctly retained abstention. This is a localized regime error, not a global misfit. **d**, The v2 descriptor-based regime gate that separates the exact-balance abstention regime from active fields. The gate uses prespecified descriptors only; K , time and source-encoding labels were not used. Panel **c** is the stay check fitted excluding the replay rows; the frozen prospective scoring set is defined in the Supplementary Methods.

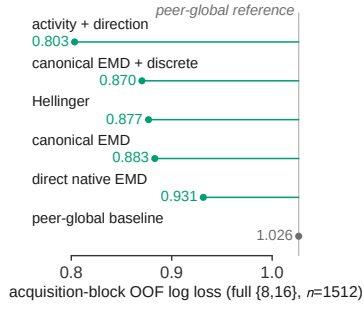


Supplementary Figure S23 (page 2 of 2). (continued: prospective validation). Prospective v1→v2 comparison over 44 new LLM collective runs (55,600 valid calls; new couplings, $N \in \{9, 17\}$, both coupling signs), on the endpoints that carry an LLM ground truth. **e**, Absolute activity error per run: v2 improves on v1 by 0.057 (95% CI 0.035–0.080). **f**, Stay-probability error: improvement 0.061 (0.037–0.084). **g**, Final- r_1 error: improvement 0.013 (0.004–0.023). **h**, Applicability-risk calibration: the prespecified applicability-risk score plotted against observed response error. Panels **e–g** are absolute errors of the compressed surrogate against the LLM-driven collective runs (44 runs, 11 cells); green arrows mark cells where v2 moved closer to the LLM and red arrows cells where it did not. Scope: the moments branch is a prospective repair of a localized regime error, not a complete predictive theory of synchronization; v2 was not uniformly superior on every endpoint.

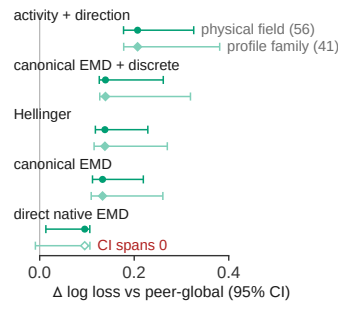


Supplementary Figure S24. Centers: in-distribution compressibility without closed-loop support. The centers surrogate predicted withheld controlled fields, but no tested collective-like field had adequate local support for prospective use. Complete evidence for stopping the centers collective-surrogate branch. The training pool is 2,736 centers rows (24 native and 15 common features), and panels **d–g** use the 36 fixed CENT-3 fields. **a**, Grouped out-of-fold log loss of the structured model versus the peer-specific global baseline, preserving physical-field grouping. The schemes shown are profile, sparse-realization, offset-group, acquisition-block and source-family holdout, with grey rows the unstructured baselines. **b**, Field-cluster bootstrap distribution and confidence interval of the log-loss improvement; acquisition-group clustering (2 clusters) has no estimable confidence interval and is omitted here. **c**, Where in-distribution prediction succeeds: activity direction, signed bias and stay probability components. **d**, Predicted versus observed response distributions on the prespecified collective-like transfer panel. **e**, Finite-peer abstention mismatch: for representative collective-like fields, the observed trinomial, the structured-model prediction and the nearest training responses; nearby active training fields did not reproduce the collective-field stay collapse. **f**, Nearest-neighbour response panel over the 36 diagnosis fields: neighbour distance, neighbour label dispersion and observed response. **g**, Local-support decomposition (active feature gap, locally noncompressible, supported); the final diagnosis contained no fully supported fields in the prespecified transfer panel. An active feature gap means that near training neighbours exist but are active where the field abstains; locally noncompressible means that the neighbourhood's own responses disagree. **h**, Peer-16-only robustness of the in-distribution and transfer checks; the ordering inverts between the two checks, the global model being worst in-distribution and best on transfer. **i**, The prespecified stop decision: STOP (microscopic archive); no confirmatory collective acquisition was run.

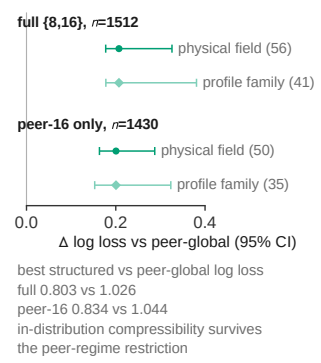
a Structured vs peer-global OOF



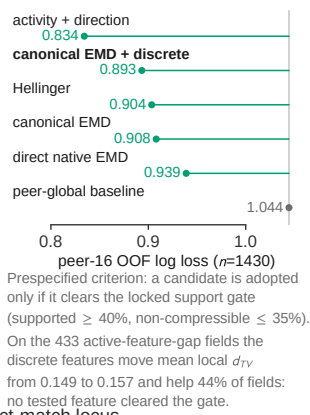
b Cluster-bootstrap improvement



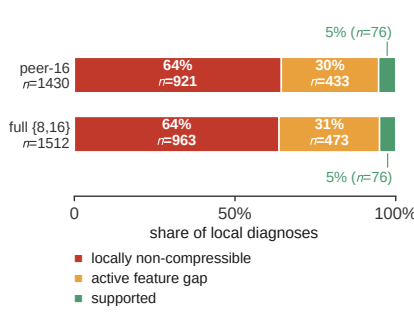
c Full vs peer-16 only



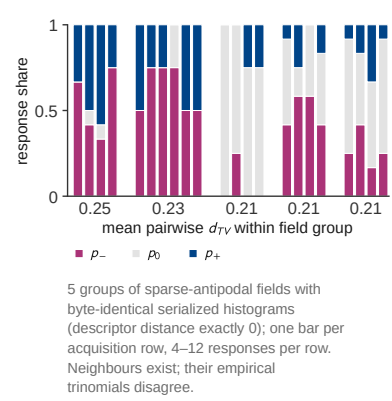
d Discrete-feature bake-off



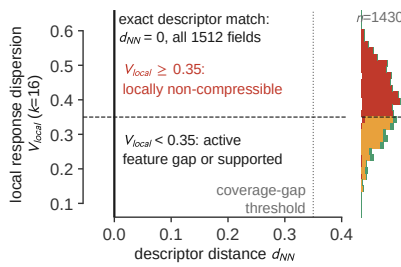
e Local-support decomposition



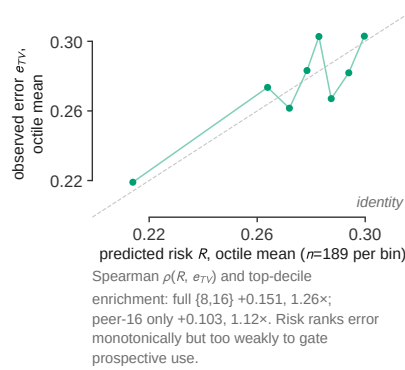
f Sparse-antipodal disagreement



g Exact-match locus



h Applicability-risk calibration



i Prespecified stop decision

active feature gap	30.3% $\leq 35\%$	PASS
non-compressible	64.4% $\leq 35\%$	FAIL
gap + non-compressible	94.7% $\leq 45\%$	FAIL
supported	5.3% $\geq 40\%$	FAIL
↓		
model gate (OOF)		PASS
support gate		FAIL
INT-3 prospective pilot		NO-GO
collective bundle freeze		no
paid authorization		no
STOP		
available finite-peer manifold does not support reliable prospective transport under the tested descriptors and coverage		

Prespecified diagnosis rule: $d_{NN} > 0.35$ (no neighbour); else $V_{local} \geq 0.35$ (locally non-compressible); else high-stay fraction < 0.25 with active fraction ≥ 0.5 (active feature gap); else supported.

Supplementary Figure S25. Intervals: matching descriptors did not yield consistent responses. Exact descriptor matches existed, but the language-model responses at those matches were too inconsistent to support prospective use. Why the intervals collective-surrogate branch was stopped despite positive in-distribution performance. Panels **a–c** use the full collective scope ($n = 1,512$; peer $\{8, 16\}$), panels **d, f, g** and **i** the peer-16 block ($n = 1,430$), panel **e** shows both, and panel **h** the full-scope risk calibrator. **a**, Structured versus peer-global grouped out-of-fold log loss. **b**, Field-cluster bootstrap distribution of the prespecified improvement statistic. **c**, Full-data versus peer-16-only analysis: in-distribution compressibility persists after restricting the peer regime. **d**, Discrete-feature bake-off: candidate interval-specific descriptor additions under the prespecified selection criterion; no tested feature made the branch eligible for prospective evaluation. **e**, Local-support decomposition among the 1,430 peer-16 collective-like fields: 64.4% locally noncompressible, 30.3% active feature gap, 5.3% supported. **f**, Sparse-antipodal disagreement: fields with close descriptor-space neighbours whose empirical trinomials diverge. **g**, Exact-match locus. The nearest-neighbour distance is exactly zero for all 1,512 fields (every field has an exact descriptor match), so $d_{NN} \equiv 0$ is drawn as a line rather than a scattered coordinate, with the local label dispersion V_{local} on the y axis, stacked by diagnosis. The coverage criterion therefore cannot fire, so the decision reduces to the V_{local} line and the strip to the right of **g**: the stop is driven by local label dispersion (response distributions disagreeing at the same descriptor position), not by missing nearest-neighbour support. **h**, Applicability-risk calibration: risk plotted against observed error. **i**, The prespecified stop decision: STOP, because the available finite-peer domain does not give adequate closed-loop support for a prospective evaluation. *Locally noncompressible* is a statement about this descriptor set and this acquisition, not a claim that no fixed response law exists. The dispersion is consistent with missing descriptors, finite-sample noise, backend drift, provenance aliasing, or a genuinely non-field-functional response; the analysis does not separate these, so the stop is reported as a decision under the prespecified rule rather than as evidence for any one account.

Supplementary Table S1: Serialized observation lengths and provider token counts

Supplementary Table S1. Serialized observation lengths and provider token counts. Fixed-width serialization makes character and token counts constant within each encoder form; both are verified against the stored per-call usage of the stored acquisitions. The intervals encoder has separate collective (finite-peer) and dense 240-peer forms; the dense form was used only in the GPT microscopic sweeps, so the Anthropic and Google backends (replay panel and, for Anthropic, the Claude collective sessions) only ever received the collective forms. Retry accounting is covered by Supplementary Table S2.

Observation map (encoder form)	Characters	Input tokens per call		
		OpenAI	Anthropic	Google
Moments (all acquisitions)	789	200	214	230
Centers (all acquisitions)	1,429	512	550	733
Intervals, collective form	1,562	609	671	925
Intervals, dense 240-peer form	1,537	604	n/a	n/a

Supplementary Table S2: Acquisition parameters and response-parser disposition

Supplementary Table S2. Fixed model identifiers, generation parameters and the disposition of every model call. Values are read from the stored per-run `resolved_config.json` and `protocol.json` artifacts. The temperature is recorded in the resolved configuration of every microscopic acquisition; the macroscopic collective runners do not echo it, and the value given is the runner default. Sampling parameters not listed (nucleus/top- p in particular) were not set in the API request, so the provider default applied and the numeric value was not recorded. Each backend calls its provider’s official Python SDK at the default public endpoint; no custom base URL was set, so no self-hosted or proxied endpoint was queried. All acquisitions fall between 2026-07-23 and 2026-07-30. Seeds were derived deterministically by the orchestration layer, but only the OpenAI and Google calls transmit one; the Anthropic Messages API accepts no request seed, so for Claude reproducibility rests on the recorded responses.

Family	API backend	Fixed model ID	Temp.	Max output tokens
GPT	openai	gpt-5.4-mini	0.7	20 [†]
Claude	anthropic	claude-haiku-4-5-20251001	0.7	20
Gemini	google	gemini-3.5-flash	0.7	20

[†]One GPT diagnostic acquisition used a maximum of 16 output tokens; all others used 20. Maximum three attempts per call throughout.

Which parser stage read each call, by acquisition:

Acquisition	Strict	Rescued	Invalid	Response form
GPT, collective runs	122,400	0	0	bare JSON
GPT, microscopic acquisitions	166,007	0	73 [§]	bare JSON
Claude macro (main session)	0	132,651	17 [¶]	fence-wrapped JSON
Claude macro ($K = 0$ control)	5,100	0	0	bare JSON
Claude replay	0	2,304	0	fence-wrapped JSON
Gemini replay	1,514	767	23	mixed

[§]All 73 occurred in the `centers_cent3` diagnostic acquisition, the one run at a 16-token output cap, where truncation broke strict JSON; they were excluded, never coerced. Every other GPT microscopic call parsed strictly. [¶]Invalid attempts in aborted rounds; none entered a retained trajectory. All counts are recomputable from the released `runs/**/trace.jsonl`.

The Gemini rescue count also depends on the encoding: 170 for moments, 275 for centers and 322 for intervals. Invalidity is not concentrated in any encoding; the rate of strict-contract compliance is. The two collective runners handle an exhausted retry differently: the Claude runner aborts the round before the agents are updated, so no partially updated round can enter the analysis, whereas the GPT runner applies a null action to the affected agent and continues, gated by a minimum valid rate of 0.95 per acquisition. That path was never taken in the retained runs, in which every call was valid.

Supplementary Methods: Fourier fitting conventions

These are the implementation details of the harmonic decomposition summarised in the main Methods.

Coefficient convention. In $g_{\mathcal{R}}(\delta) = a_0 + \sum_m [a_m \sin(m\delta) + b_m \cos(m\delta)]$, a_m multiplies the sine. This is the reverse of the convention in which a_m is the cosine coefficient, matching the stored artifact order; a_m , not b_m , is the odd component.

Averaging across blocks. The fit is carried out separately within each acquisition block and the resulting complex coefficients $C_m = a_m + ib_m$ are averaged across blocks before conversion to magnitude $R_m = |C_m|$ and phase $\phi_m = \arg C_m = \arg(a_m + ib_m)$, so phases are never averaged directly. The least-squares fit is unweighted, giving the 36 offsets equal weight irrespective of how many valid

responses each contributed.

Bootstrap and ellipses. Uncertainty on (a_m, b_m) comes from a multinomial bootstrap within each offset, holding the total number of observations at that offset fixed and redrawing the trinomial counts from the observed proportions. The number of resamples was set per acquisition: 200 for the concentration sweep, 100 for the stimulus-domain sweep (artifact name `stimulus_manifold`) and 2,000 for the serialization feature sweep. The 95% region plotted for a coefficient is a chi-square ellipse on two degrees of freedom at the quantile 5.9915 (the 95% point of χ_2^2), constructed from the sample covariance (denominator $n - 1$) of the bootstrap draws.

Phase display gate. The phase of a complex coefficient is unstable and effectively undefined when its magnitude is near zero, so a phase is reported only when the amplitude is large enough and its uncertainty region excludes the origin. Within an acquisition block, both $m = 1$ and $m = 2$ had to satisfy $R_m \geq 0.15$ with an ellipse excluding the origin. For the block-averaged coefficients the ellipse condition was applied to $m = 1$ only, in the conservative form that the phase is suppressed if the ellipse of either block covers the origin, whereas $m = 2$ was gated on $R_2 \geq 0.15$ alone. The value 0.15 is a visualization and reporting convention, not an inferential endpoint.

Supplementary Methods: surrogate compressibility and transportability

This section specifies the surrogate-modelling and support-diagnosis procedures summarised in Supplementary Figures S21–S25.

Notation

A field ρ enters the surrogate only through its prespecified descriptor vector $x(\rho)$. These descriptors describe the physical field itself and are computed from the same 24-bin record in every branch (Feature sets below); they are not the prompt any encoding presented. The measured quantity is the language model’s empirical action distribution $\hat{\mathbf{p}}(\rho) = (\hat{p}_-, \hat{p}_0, \hat{p}_+)$, estimated from $n(\rho)$ repeated model calls with action counts $c_f(\rho)$, $f \in \{-1, 0, +1\}$. The predicted quantity is $\mathbf{q}_\theta(x(\rho))$, a distribution over the same three actions. Fitting and every reported comparison use the multinomial log loss, computed over the set of measured fields $\rho_1, \dots, \rho_M$ entering the fit or the evaluation,

$$\ell = - \frac{\sum_{i=1}^M \sum_{f \in \{-1, 0, +1\}} c_f(\rho_i) \log q_{\theta, f}(x(\rho_i))}{\sum_{i=1}^M n(\rho_i)}, \quad (15)$$

where θ collects the parameters of whichever candidate model is being fitted; ℓ is the mean negative log-likelihood per model call in nats, and it shrinks exactly when the model places higher probability on the actions that actually occurred. A reported improvement is $\Delta\ell = \ell_{\text{baseline}} - \ell_{\text{model}}$, with the peer-count baseline $\bar{\mathbf{q}}^{(m)}$ of the Baseline subsection as the reference. The support diagnosis uses two field-level statistics defined under Local-support diagnosis: $d_{\text{NN}}(\rho^*)$, the distance from a closed-loop field ρ^* to the most similar training field, and $V_{\text{local}}(\rho^*)$, which measures how strongly the measured responses of its $k = 16$ nearest training rows disagree with one another. Prospective evaluation compares per-run absolute endpoint errors $e = |E_{\text{pred}} - E_{\text{obs}}|$; the v1–v2 comparison below reports

$$\Delta e = e_{v1} - e_{v2}.$$

Feature sets

For each observation map the surrogate consumes only the prespecified descriptor vector derived from the 24-bin field; K , absolute phase, agent identity, time step and source-encoding labels are excluded from every feature set. The moments branch uses the 15 descriptors common to all three encodings: the real part, imaginary part and modulus of each binned circular moment $\tilde{z}_1, \tilde{z}_2, \tilde{z}_3$ (nine values), together with circular entropy, antipodal balance, asymmetry, bimodality, sparsity and the peer count (defined as `COMMON_FEATURE_NAMES` in `circlemap/field_features.py`). The centers branch receives the same 15 descriptors followed by the 24 normalized bin masses in canonical order, 39 inputs in total, and the intervals branch receives exactly the same 39-dimensional vector. The two histogram branches therefore present the surrogate with the same feature vector for the same field, which is what makes any difference in their measured responses a serialization effect rather than a feature-space effect.

Candidate model classes and hyperparameters

The response is parameterised as the trinomial action distribution (p_-, p_0, p_+) over $\{-1, 0, +1\}$. The candidate model classes were: a constant empirical trinomial; a peer-count-stratified constant trinomial; a kernel nearest-neighbour predictor; multinomial logistic regression with an L_2 penalty; boosted softmax stumps; a hurdle multinomial, which first models whether the agent acts at all and then, given action, its direction; and kernel and regime-aware kernel variants of that hurdle. Hyperparameters were fixed at the values written into the model definitions rather than searched, so there was no tuning loop and no held-out selection over hyperparameters. The values are: L_2 penalty 1.0, learning rate 0.05 and 250 gradient steps for the logistic and hurdle models; 40 estimators, learning rate 0.35 and 8 candidate thresholds for the boosted stumps; length scale 1.5 with $k = 32$ for the kernel neighbour predictor; and length scales 1.2 (stay) and 1.5 (direction) with $k = 32$ and mixing weight 0.5 for the kernel hurdle.

Grouped cross-validation split

A row of the modelling table is one (encoding, profile, offset index, acquisition block) cell. Because every physical field, that is every (profile, offset index) pair, was measured twice in two acquisition blocks, each physical field contributes two rows. Cross-validation then proceeds as in the main text: the rows are split into folds, each fold is hidden in turn, the model is fitted on the remaining rows, and the loss is evaluated on the hidden fold only. Five ways of forming the folds are evaluated. Four of them keep both rows of a physical field in the same fold, so that no field is split across train and test: `leave_one_profile_out` (41 folds), `offset_group_holdout` (6 folds), `sparse_realization_holdout` (6 folds) and `source_family_holdout` (3 folds). The fifth, `acquisition_block_holdout` (2 folds), splits by acquisition block and therefore deliberately places the two rows of the same physical field on opposite sides: because the same field then appears in training and in test, this scheme measures whether a fitted response model reproduces across acquisition blocks, not whether it extrapolates to fields never seen in training, and it serves as the primary model-selection criterion. In every scheme the folds are simply the distinct values of the grouping label, each held out in turn; the assignment is deterministic, involves no shuffling and therefore needs no fold-assignment seed. Where a grouping label is undefined on some rows, as for `sparse_realization_holdout`, only the labelled rows are evaluated. The reported loss is the multinomial log loss of the Notation subsection, weighted by the response counts of each row. One reporting detail matters when comparing tables: an out-of-fold loss

can be aggregated either by pooling all held-out rows and computing one loss over the pool, or by computing the loss within each fold and then averaging the folds, and with unequal fold sizes the two give slightly different numbers. The centers and intervals branch reports use the pooled form, the scheme-by-model comparison uses the fold mean, and each table states which form it shows.

Baseline

The comparison baseline deliberately ignores the shape of the field. For a given peer count it always predicts the same three action probabilities, namely the average response observed over all training fields with that peer count. Beating it therefore requires using information about the field beyond how many peers it contains. We call this the peer-specific global baseline. Concretely, the baseline keeps one constant trinomial per peer count: it returns the marginal (retard, stay, advance) frequency of the training rows carrying that peer count, after adding a Laplace smoothing constant of +0.5 to each of the three cells and renormalizing (the smoothing only prevents zero probabilities). No field descriptor of any kind enters the baseline, and there is no additional conditioning on coupling, time, source encoding or acquisition block. A peer count never seen in training falls back to the pooled training distribution over all peer counts. In the collective scope the peer counts are $\{8, 16\}$.

Model-selection criterion

The model class was selected by the grouped out-of-fold log loss under the acquisition-block holdout, judged relative to the peer-specific global baseline. A candidate first had to beat the constant empirical baseline to be eligible at all. Among eligible candidates, any within 2% of the best loss was treated as tied, and ties were broken by a fixed simplicity preference, in the order kernel nearest neighbour, hurdle multinomial, multinomial logistic, boosted stumps. If nothing beat the baseline, the lowest-loss candidate was reported without being adopted.

Resampling unit for the surrogate

Repeated model calls were never the bootstrap unit. The cluster for the surrogate log-loss comparison is the *stimulus profile*, the identifier that aggregates a stimulus across its rotation offsets, of which there are 56. This is coarser than the *physical field* of the replay and presentation experiments, the profile-and-offset unit, of which this dataset contains 756. We reserve *physical field* for the finer unit throughout and name the surrogate resampling unit the stimulus profile wherever a surrogate interval is reported. Intervals on the prospective v1–v2 comparison are a different resampling again: there the unit is the collective run, with 2,000 run-level percentile-bootstrap resamples.

The two moments surrogate versions

Version v1 is the response model fitted to the microscopic sweeps and frozen before any collective field was scored; it was retained unchanged as the prospective baseline, so the comparison below is not a refit against its own test set. Checking v1 against the measured replay responses on collective fields exposed one systematic failure and one preserved success. The failure was that it overpredicts **stay** on active collective fields, that is, it predicts abstention where the model in fact acts. The preserved success was at exact balance: on fields with $\varepsilon = 0$, where the peers are symmetric about the focal agent and the measured operator does abstain, v1 continued to predict abstention, so these fields act as

anchors that a revision must not break. Version v2 adds the regime gate defined in the next subsection, computed from the prespecified field descriptors alone.

Prospective validation comprised 44 language-model runs and 55,600 valid calls across new coupling values, $T = 100$, $N \in \{9, 17\}$ and both coupling signs. Improvement was defined per run as

$$\Delta e = e_{v1} - e_{v2}, \quad (16)$$

where e_{v1} and e_{v2} are the absolute errors of the two versions on the same endpoint of the same run, so positive values favour v2. The frozen prospective scoring set comprises activity, stay probability and final r_1 ; social torque and terminal-lock time were not included in it and are therefore not reported as prospective error endpoints, which does not imply they are unobservable from the language-model runs.

v2 stay-gate (descriptor-only regime gate)

The v2 moments surrogate adds a regime gate computed from the prespecified field descriptors alone (with the same exclusions as in Feature sets), so that exact-balance abstention fields are routed separately from active fields. The gate reads two of the common descriptors, antipodal balance and $|\tilde{z}_1|$, and routes a field to the abstention expert when antipodal balance ≥ 0.95 and $|\tilde{z}_1| \leq 0.12$, and to the active expert otherwise. The two experts are kernel neighbour predictors that differ in length scale and neighbourhood size: 1.2 with $k = 24$ for the abstention regime and 0.25 with $k = 6$ for the active one.

Local-support diagnosis

The neighbour search is carried out in the encoding-native 24-bin mass vector, not in the common descriptor vector used elsewhere, and no per-feature scaling or standardisation is applied. The distance is the circular earth-mover distance between two 24-bin mass vectors: picturing each histogram as sand distributed around a circle, it measures how far the sand must be carried around the circle to turn one histogram into the other. Operationally, both vectors are normalised to sum to one, the cumulative sum of their difference is taken around the bins, that cumulative sum is centered by subtracting its own mean, and the distance is the mean of the absolute values of the centered cumulative sum over the 24 bins; $d_{NN}(\rho^*)$ is this distance from ρ^* to its nearest eligible training row. Only training rows with the same peer count as the query row are eligible as neighbours, and the neighbourhood comprises the $k = 16$ nearest training rows (fewer when fewer eligible rows exist). The label-dispersion statistic V_{local} is the mean total-variation distance over all pairs of those k neighbours' empirical trinomial action distributions, that is over the 120 distinct pairs when $k = 16$; it is large exactly when physically similar fields drew visibly different responses from the language model.

Category thresholds

Each diagnosed field is assigned to exactly one of the three categories used in Supplementary Figures S24g and S25e. A coverage gap is declared at $d_{NN} > 0.35$ in both branches, but the condition differs: the intervals branch uses $d_{NN} > 0.35$ on its own, whereas the centers branch requires $d_{NN} > 0.35$ together with fewer than five effective neighbours. `local_noncompressible` is declared at $V_{\text{local}} \geq 0.35$ in both branches; the centers branch carries a second, lower dispersion threshold of 0.25, used to separate its stay-model-bias category. A field is **supported** when d_{NN} is within support, V_{local} is below the

applicable dispersion threshold, and a sufficient fraction of the neighbours is high-stay; it is an active feature gap (**feature_gap**) when d_{NN} is within support but the neighbourhood is predominantly active while few of its neighbours are high-stay. The neighbour-level stay and activity cutoffs and the required neighbour fractions entering these two composite categories are read from the locked support-gate constants in the released manifest.

Decision order. The categories are mutually exclusive because they are evaluated as an ordered cascade: the first condition that fires assigns the label and the remaining conditions are not consulted. There is no separate precedence table; the order in the code *is* the precedence rule, and it differs between the two branches:

Branch	First test	Second test	Third test
intervals	coverage gap	local dispersion	active feature gap
centers	coverage gap	active feature gap	local dispersion

A field reaching the end of the cascade without firing any condition is labelled **supported**; the branch-specific conditions are those fixed in Category thresholds above.

The order matters in principle, because a field could satisfy both the dispersion and the active-feature-gap conditions and would then receive different labels under the two orders. It does not change either stop decision here: for centers the diagnosis contained no supported fields under either order, and for intervals the coverage condition cannot fire at all, because d_{NN} is exactly zero for every field in both stored support tables (Supplementary Fig. S25g).

Reproducibility and decision audit

This section records what was hash-locked and when, in what order the branch decisions were frozen, and what software produced the numbers, for checking the reported analyses against the released repository.

Prespecification and acquisition integrity

Throughout, an *artifact* is a single file written by the pipeline and never edited afterwards: a protocol or configuration file, a serialized prompt set, a raw run directory, or a table of computed statistics. Protocols, observation maps, prompt text, model IDs, physical seeds, analysis endpoints and stopping rules were versioned and SHA-256 locked before the corresponding acquisitions, and the locks were verified by the runners at acquisition time. We use *prespecified* throughout in this internal sense: these records were written, versioned and hash-locked in the project repository before the acquisition or analysis they govern, but they were not deposited with a public registry, so the ordering is documented by our own records rather than externally attested. For the identical-field replay the guarantee is also structural (Supplementary Fig. S11).

Raw responses, parser status, retry history, serialized observations, actions and phase trajectories were retained. No failed response was silently converted into a valid action.

The repository carries the operational record that this section summarizes: the machine-readable hash manifest (individual hash values and artifact paths), the individual request URLs, the SDK patch levels where they were recorded, and the complete per-call retry history behind Supplementary Table S2. It also carries the acquisition traces themselves, one record per model call across all 15 acquisition groups, so every count reported here can be recomputed from the raw responses rather

than taken on trust.

Stop-or-repair chronology

The three branch outcomes (moments prospective repair, centers STOP with microscopic archive, and intervals STOP) followed a prespecified chronology: descriptor-space coverage check, in-distribution grouped-CV compressibility test, collective-like transfer test, local-support diagnosis, and the prespecified stop-or-repair decision. The table below gives the time at which each step’s artifact was written, together with the artifact that carries it, so that the ordering can be checked directly against the released tree. Times are UTC. The two STOP rows are the closure times recorded inside the manifests themselves (`closed_at_utc`); the remaining rows are the write times of the named artifacts.

Step	Moments (repair)	Centers (STOP)	Intervals (STOP)
Descriptor-space coverage	2026-07-23 15:33	2026-07-24 04:21	2026-07-24 05:58
In-distribution grouped CV	2026-07-23 15:53	2026-07-24 04:22	2026-07-24 05:59
Collective-like transfer	2026-07-23 23:45	2026-07-24 04:47	not acquired [‡]
Local-support diagnosis	2026-07-23 23:49	2026-07-24 05:41	2026-07-24 06:07
Stop-or-repair freeze	2026-07-24 01:05	2026-07-24 05:58:41	2026-07-24 06:07:23

The artifact paths for each row are listed in the repository manifest. [‡]The intervals branch stopped before any prospective transfer acquisition was authorized, so no paid transfer acquisition exists for that branch.

Software and library versions

- Language and runtime: CPython 3.11 on Windows. All local random draws in the simulation and statistical analyses use explicitly seeded `numpy` generators; provider-side language-model sampling is governed separately by the API settings reported in Supplementary Table S2.
- Numerical and statistical libraries: the analyses import `numpy`, `pandas` and `matplotlib` and nothing else. Every permutation test, bootstrap, exact sign test and rank correlation is implemented directly against `numpy`, so no result depends on the defaults of a statistics package; neither `scipy` nor `scikit-learn` is a dependency. The project environment carries `numpy` 2.4.6 and `pandas` 3.0.3; the figures were rendered with `matplotlib` 3.10.9.