FROM SYSTEM SYMPTOM TO AGENT-HISTORY DIAGNOSIS: A THREE-COORDINATE PROTOCOL FOR LLM POPULATION DRIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Behavioural change in multi-agent LLM populations may reflect shared time, heterogeneous agent histories, or visible interaction evidence. We introduce DriftClock, a three-coordinate protocol separating shared time, private exposure, and visible history. In naming games, round and exposure correlate at 0.91; exact same-round contrasts instead localize the signal to agent histories: switching after disagreement declines with exposure, future exposure is null, a 64-population Mistral study replicates exposure-indexed dynamics, and long-horizon decline transfers across models and objectives. Matched $H \in \{10, 30, 60\}$ memory-window interventions reveal a finite-context breakpoint: response falls as histories fill and partially rebounds when old evidence leaves longer windows. Discovered in Qwen, this breakpoint passes a frozen Mistral $H30 - H60$ test and an independent-seed Qwen replication; social-update rules separately show that exposure signatures need not imply autonomous ageing. Standardized challenges preserve eventual adaptation, while post-hoc common-support analyses locate the raw speed gap in measured visible evidence. Paired history branches separate local source effects from propagation: history differences reach other agents while action differences remain sparse and model-dependent. DriftClock moves from population symptom to history localization, formation, functional consequence, and transmission, establishing finite-history turnover as a reproducible boundary condition on cumulative-age explanations.

1 INTRODUCTION

End-task evaluation says whether a multi-agent system failed; recent taxonomies and attribution benchmarks ask which agent and step were responsible (Cemri et al., 2025; Zhang et al., 2025; Chen et al., 2026). Yet gradual behavioural change can emerge before a terminal failure, and an aggregate trajectory does not reveal whether the population changed together or particular agents accumulated different experiences. Conventional monitoring plots such change against episode length or population round. That axis is not neutral: when only a subset acts each round, agents sharing the same calendar time have different local histories. We use the contrast between the shared population clock and each agent's exposure clock as a localization probe, then inspect the visible histories carried by the implicated agent-history units.

We study this problem in a naming game, a minimal coordination task in which agents form a shared convention (Baronchelli et al., 2006; Ashery et al., 2025). It isolates local choices, feedback, and collective state without a complex task pipeline. We ask three diagnostic questions: **which agent-history states carry the change, what kind of change do they carry, and does it transmit downstream?** The target is stateful: the same agent may occupy different diagnostic states over time. In natural trajectories, *plasticity* is the probability of changing name after the agent's previously recorded partner disagreed. We call its exposure-linked reduction a *plasticity decline*; it may stabilize a useful convention and does not imply an internal age variable or permanent incapacity to update.

The two-clock design separates two time indices while tracking collective state:

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

- **Individual exposure** $e_i(t)$ counts agent i's prior interactions; it indexes, but is not identical to, the length or content of its visible history.
- **Shared time** t counts population rounds. Population state $s(t)$, the share using each convention, describes the evolving collective environment rather than another age count.

Architectural anchor. Frozen weights and no state outside the prompt imply identical decision distributions for identical tokenized inputs under the same decoder. Exposure can enter through rendered indices or interaction-produced history; removing indices leaves the latter channel. With a finite window, exposure lengthens visible history until $e = H$ and thereafter replaces old evidence, supplying a testable change point. Individual history is itself socially generated, so a history-mediated explanation need not exclude collective feedback.

In our logs exposure and round correlate at 0.913. We compare agents in the exact same population and round, fixing shared time and pre-decision collective state, and cluster-bootstrap population runs. The comparison identifies an exposure/history association, not a count effect with history content fixed. Wider time cells, joint state models, a future-exposure placebo, and population-size scaling provide complementary diagnostics. We separately introduce a standardized sustained challenge because natural disagreement does not hold the prior evidence faced by agents fixed.

Paper map and findings. We apply DriftClock in two regimes: free convergence and pressure from a 12.5% committed minority. Its three coordinates organize five linked acts: I separates shared time t from exposure e; II localizes the e-indexed signal; III inspects visible history/state z; IV fixes the future challenge; and V branches one source history to measure transmission:

1. **Population symptom.** Round and exposure correlate at 0.91, so the usual population-time curve mixes shared duration with unequal individual histories.
2. **Agent-history localization.** Plasticity falls from 0.27–0.30 in the first five interactions to about 0.02 after thirty. Exact same-round slopes are negative, future exposure is null, and an index-free cross-size test aligns curves more closely in exposure units.
3. **Formation pathway.** Visible-history edits and social-update rules show that accumulated history can generate the localized signal without autonomous ageing. Matched window interventions expose a filling-to-turnover breakpoint across Qwen and Mistral, while paired reordering is model-conditional.
4. **Functional diagnosis.** Every challenged snapshot eventually changes. Higher-exposure histories respond more slowly in the raw comparison, and four common-support specifications locate that gap in measured visible evidence.
5. **Source and transmission diagnosis.** Paired branches localize a strong causal response to one source history, then reveal that history differences spread much more readily than action differences; repeated contact produces small, delayed downstream leakage.

Contribution. We contribute a three-coordinate diagnostic that moves from a system-level symptom to an agent-history unit, a supported formation pathway, and a paired propagation test. Unlike responsibility attribution after a failed task, DriftClock localizes a behavioural phenotype and the histories that carry it: exposure is a diagnostic coordinate rather than a randomized treatment. Window-length interventions turn the architectural boundary $e = H$ into a prospective formation test; paired source branches then distinguish local history sensitivity, downstream memory reach, and action divergence. Across the sequence, the protocol yields evidence grades—round-confounded, exposure-indexed, content-mediated, source-localized, transmission-attenuated, or unresolved—and distinguishes delayed, absent, and preserved adaptation.

2 RELATED WORK

Failure localization and longitudinal drift. MAST categorizes failed multi-agent traces into system-design, inter-agent, and task-verification failures (Cemri et al., 2025). Who&When links failures to responsible agents and decisive steps (Zhang et al., 2025); TraceElephant shows that full execution traces improve such attribution (Chen et al., 2026). These works diagnose completed task failures. DriftClock instead localizes a gradual response phenotype before assuming a terminal failure: the unit is an agent with its accumulated history, not a blame label. Adjacent work documents

instruction instability (Li et al., 2024), long-horizon coherence failures (Backlund and Petersson, 2025), reliance on early answers (Laban et al., 2026), conversational inertia (Wan et al., 2026), and action-history bias (Zhang et al., 2026). Al-Tawaha et al. (2026) use fixed probes and order randomization to separate memory content from encounter order. These findings motivate history accumulation as a candidate pathway, not an established mechanism in our task.

Coordination, memory, and collective persistence. Naming games (Baronchelli et al., 2006) and committed-minority experiments (Centola et al., 2018; Ashery et al., 2025) make convention formation and change measurable. Ashery et al. (2025) already connect memory-dependent individual choices to collective bias; history dependence itself is not our novelty. Tanaka (2026) model mutual in-context learning in naming games and derive population-level scaling laws for consensus. Mehdizadeh and Hilbert (2026) show that memory and topology interact, with faster settling sometimes reflecting fragmentation rather than consensus. De Marzo et al. (2026) obtain collective persistence without individual memory or weight updates, while Okawa (2026) study biased consensus in debates. Individual history and collective feedback are therefore competing and potentially coexisting channels. Our addition is an exposure-conditioned account of individual plasticity during coordination, not a claim that collective dynamics reduce to independent agents. Local-history models (Guilbeault et al., 2026) motivate transparent social-update references.

Attribution and validity. Observational-equivalence critiques (Barrie and Törnberg, 2025) and consensus-validity diagnostics (Yang, 2026) caution against reading mechanisms directly from aggregate trajectories. Existing work separately studies longitudinal memory risk and population-level consensus mechanisms. DriftClock targets the attribution confound among shared time, individual exposure, and visible evidence in asynchronous populations, and organizes their separation as a sequential diagnostic protocol. Same-round comparisons isolate the exposure-indexed signature from shared time; social baselines, history normalization, and a common challenge then test its specificity and behavioural meaning.

3 TESTBED AND MEASUREMENT

$N = 24$ agents per population, instantiated from Qwen2.5-7B-Instruct, play a naming game over a calibrated symmetric two-name pool: each round one pair is drawn (rotating neighbourhoods of size 5, reassigned every 25 rounds); both agents independently choose a name given a rendered window of their last $H = 30$ interactions; matches are rewarded, mismatches penalised, and outcomes appended to both memories. Decisions are single-token argmaxes, so each is one forward pass. Two regimes, 32 populations each, 400 rounds: *free convergence* (populations settle onto a convention endogenously; 24,832 usable decisions) and *committed minority* (12.5% of agents always emit the competing name; 21,671 usable decisions from the non-committed agents).

The two clocks in the logs. For every decision we log the calendar round t, the population state $s(t)$, and the acting agent's private exposure e (its interaction count so far). One pair acts per round. The first endpoint is uniform and the second is sampled from its currently assigned random neighbourhood, which is redrawn every 25 rounds. Marginal participation is therefore approximately $1/12$ per round, but directed neighbourhoods create temporary selection-probability differences; the observed within-round exposure spread ($\mathrm{sd} \approx 5.5$ at $t = 400$) powers the comparison.

Outcome variable. Plasticity must be defined with care, because the obvious definition is a collider: $P(\text{switch} \mid \text{THIS round's partner disagreed})$ conditions on an outcome of the same round—switching *to match* the partner mechanically lands in the conditioning set. We therefore use the agent-observable version, $P(\text{switch} \mid \text{the agent's most recently recorded partner disagreed})$, which conditions only on information available before the decision; the raw switch rate and the collider version are reported alongside and agree in shape (Appendix A).

Dependence and option order. We randomise option order per decision and cluster inference at the population-run level (Cameron and Miller, 2015) to account for within-population dependence. In the Stage 8A replay, the model selects the first-listed option in 0.484 of all 15,744 scored decisions (0.500 at the snapshot decision, where 0.995 of choices are the incumbent), providing an artifact-backed check that the response-area endpoint is not driven by option position.

DriftClock diagnostic protocol. Let t denote shared time, e private exposure, and $z = (s, h)$ the pre-decision population state and visible history. The diagnostic unit is an agent-history snapshot

(i, t, e, z), not a permanent agent trait. The protocol is sequential: (i) separate t and e with exact-time contrasts, placebos, and rate changes; (ii) test specificity with index-free renderers, social rules, common-support matching, and event-order interventions within h; (iii) fix the future challenge and measure response and adoption; (iv) branch a full checkpoint, edit one source history, and track downstream divergence. Table 1 states what each step licenses. A post-hoc label is provisional; an overlap null is not equivalence, and a source effect is not propagation.

Table 1: Sequential diagnostic questions and the claims their contrasts license.

diagnostic question	contrast	warranted label
Population-wide or history-localized?	same-round contrast; placebo; size scaling	round confound; exposure index
Count, content, or order?	index-free runs; social rules; visible-z matching; paired reordering	content-mediated; order-sensitive; unresolved
Adaptation delayed or absent?	fixed challenge: response area and adoption	delayed; absent; preserved
Source-localized or transmitted?	source-only paired branch; matched future schedule	source effect; attenuated or propagated

4 RESULTS

4.1 ACT I—POPULATION SYMPTOM: ONE CURVE, TWO CLOCKS

Marginally, exposure and round are nearly the same variable: $\mathrm{corr}(t, e) = 0.913$ (free) and 0.912 (committed). A round-indexed curve therefore superimposes shared duration, unequal interaction counts, and the histories those interactions produce. The first diagnostic outcome is a confound, not yet a winner: shared time must be held fixed before the individual-history signal can be read.

4.2 ACT II—LOCALIZATION: INDIVIDUAL HISTORY BEYOND SHARED TIME

Private exposure at exact collective time. Our main post-hoc estimator compares the two non-committed agents acting in the same population and round. This fixes calendar time and the pre-decision population state exactly; the remaining contrast is indexed by their exposure difference, which also bundles differences in accumulated local history. For exact run-round cells c, with agents a and b, we estimate

$$\widehat{\beta}_e = \frac{\sum_c (e_{ca} - e_{cb})(y_{ca} - y_{cb})}{\sum_c (e_{ca} - e_{cb})^2}. \tag{1}$$

Pair selection is randomized before either current action, exposure is measured before the decision, and the two outcomes cannot affect their already accumulated exposure difference. Inference resamples complete population runs, preserving serial dependence and repeated agents. Stable agent heterogeneity is addressed separately by the within-agent comparison in §4.2. Table 2 reports linear-probability slopes with percentile intervals from 600 cluster bootstraps over the 32 population runs.

Table 2: Exact same-round exposure slopes. Values are change in switch probability per additional private interaction.

regime	unconditional switch	previous-disagreement plasticity
free convergence	-0.00181 $[-0.00237, -0.00136]$	-0.00659 $[-0.00993, -0.00425]$
committed 12.5%	-0.00180 $[-0.00239, -0.00136]$	-0.00550 $[-0.01145, -0.00246]$

The raw estimates use 11,166 and 8,442 unequal-exposure pairs; the conditioned estimates use 1,568 and 813 pairs for which both agents' previous partners disagreed. Exposure slopes remain negative with intervals excluding zero when the fixed cell is widened from one to 5, 10, or 25 rounds. In joint run-by-25-round-epoch models, the standardized residual exposure coefficients for plasticity are -0.0211 $[-0.0277, -0.0160]$ and -0.0317 $[-0.0444, -0.0224]$, whereas residual round coefficients are -0.0023 $[-0.0082, +0.0033]$ and -0.0057 $[-0.0134, +0.0030]$.

Randomization and future-exposure placebo. A run-cluster sign-flip test rejects a zero, sign-symmetric exposure slope in both regimes (plasticity $p < 10^{-4}$ free; $p = 0.0033$ committed). As a negative-control predictor, we replace accumulated exposure with the number of interactions the agent will receive *after* the current decision. Future exposure is generated by subsequent randomized pair selections and cannot cause the current action. Its exact-round slopes are $+0.00036$ ($p = 0.81$) and $+0.00055$ ($p = 0.75$) for plasticity, and $+0.00034$ ($p = 0.48$) and $+0.00050$ ($p = 0.34$) for unconditional switching. The current-exposure association is therefore not mirrored by eventual activity.

Collective time and state at exact private exposure. Within exact run-by-exposure cells, residual round is positive rather than negative ($+0.00560$ $[+0.00130, +0.01129]$ free; $+0.01283$ $[+0.00495, +0.02156]$ committed), while support for the current choice predicts less switching (-0.01793 $[-0.02533, -0.01081]$; -0.03307 $[-0.04511, -0.02471]$). We use *exposure-indexed* predictively: exposure slopes are negative across exact and widened cells, residual round is not detectably negative in joint epoch models, future exposure is null, and changing population size stretches the round axis while approximately preserving the exposure curve. Exposure is not randomized and population state is not irrelevant; accumulated exposure histories carry the monotone trend more reliably than calendar age alone. This localizes a *level*—individual history—rather than a stable agent identity. Section 4.5 then branches a full checkpoint to localize one source snapshot and test its downstream transmission.

Magnitude and within-agent confirmation. Correlations understate what is happening, because switching is rare overall. The exposure-indexed plasticity curve shows the magnitude directly:

Table 3: $P(\text{switch} \mid \text{last recorded partner disagreed})$ by the agent's own exposure e. Run-clustered; bins with ≥ 40 conditioning decisions.

exposure e	1–5	5–10	10–15	15–20	20–25	25–30	30–40	40–60
free convergence	0.271	0.105	0.055	0.020	0.010	0.007	0.016	0.000
committed 12.5%	0.296	0.121	0.103	0.021	0.017	0.009	0.023	0.000

A fresh agent abandons its convention under disagreement about one time in three-and-a-half; after thirty interactions, essentially never. The committed regime—where a persistent minority keeps supplying disagreement—retains plasticity roughly twice as long through mid-exposure (0.103 vs 0.055 at $e \in [10, 15)$) but reaches the same observed floor. Persistent opposition therefore delays the decline in natural conditional switching; it does not establish permanent loss of updating.

Within-agent confirmation. Pooled early-vs-late comparisons can be composition artifacts (late strata contain different agents). Using *unconditional switching*, pairing within each (run, agent) cell—the same agent's switch rate in its own low-exposure quartile versus its high-exposure quartile—gives a mean late-minus-early change of -0.077 (768 cells; of the cells with a nonzero difference, 350 decrease vs. 10 increase) and -0.085 (672 cells; 353 vs. 3). This establishes within-agent decline for the unconditional outcome; it is distinct from the conditioned plasticity statistic in Table 3.

Cross-size diagnostic. With the standard renderer that prints private interaction indices, we ran $N \in \{12, 24, 48\}$ for 200/400/400 rounds (32 populations each). Exposure-indexed bootstrap bands overlap in 19/21 bin comparisons, whereas round-indexed half-lives stretch from 31 to 62 and 88 rounds as the per-agent scheduling rate falls with $2/N$ (Appendix E). This favours exposure as a transferable plotting coordinate, not as a mechanism-unique cause: social-update rules can evolve on the same $2t/N$ scale.

4.3 Act III—Formation: exposure indexes social history

Removing printed interaction indices leaves a $\sim$50$\times$ exposure-linked plasticity decline ($0.246 \rightarrow 0.005$) and a decisive within-agent contrast (-0.047): the association does not require an explicit age label. The index-free within-epoch gradient is less precise and a population-clock coefficient is detectable; because the renderer contrast was not estimated as a paired effect, current precision does not isolate how much the printed index amplifies the signature (Appendix D).

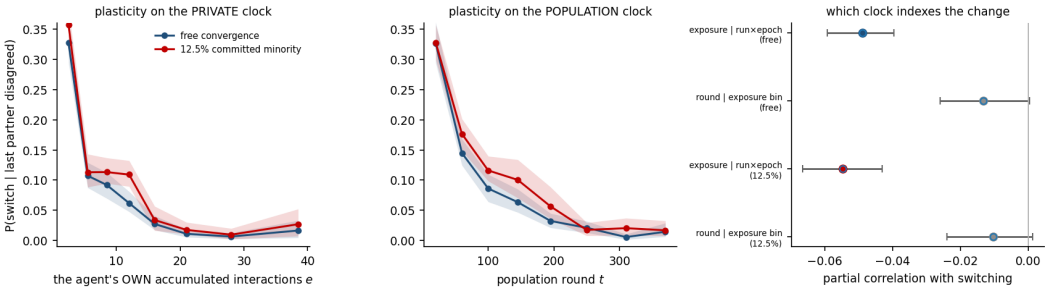

Figure 1: Left: plasticity against the agent's own exposure. Middle: the same data against calendar round—the curve a conventional monitor sees. Right: the four partial correlations; only the exposure-indexed effects exclude zero. Shaded: 95% run-level cluster bootstrap bands.

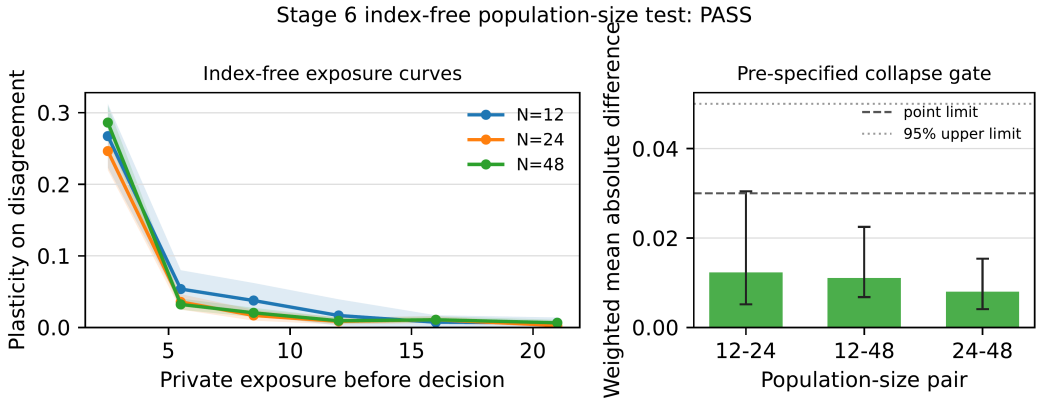

Figure 2: Pre-specified index-free size test. Left: exposure-indexed plasticity with population-cluster bootstrap bands. Right: pairwise weighted curve deviations and their 95% intervals against the fixed point (0.03) and upper-bound (0.05) margins. Green bars pass both margins.

Pre-specified index-free population-size test. We then froze a stronger collapse test before running new outputs: $N = 12, 24, 48$ with 32 populations each, horizons 200/400/800, six common exposure bins, and explicit margins. The test passes: pairwise count-weighted deviations are 0.012, 0.011, and 0.008, with bootstrap intervals [0.005, 0.030], [0.007, 0.022], and [0.004, 0.015]. These are weighted summaries, not binwise equivalence: the largest individual-bin difference is 0.0344 (Appendix E). Every size declines, and the pooled within-run slope is -0.042 per 10 interactions (95% interval [-0.046, -0.038]). A labelled exact-revision $N = 24$ rerun retains the gate. Applying the exact same-round estimator post hoc gives -0.0027 [-0.0057, -0.0001], -0.0023 [-0.0038, -0.0009], and -0.0030 [-0.0043, -0.0018] at $N = 12, 24, 48$; the archived $N = 24$ arm remains imprecise (-0.0007 [-0.0026, 0.0018]).

Transparent boundary models. Before inspecting outputs, we fixed three rules without an independent age variable: an H=30 sliding majority, the same rule with two persistence pseudo-counts, and an exponentially weighted partner-choice state (decay 0.8). Sliding majority and EWMA produce negative same-round slopes at every size. Neither establishes full collapse because late N=12 disagreement bins lack support; the inertial rule passes the pairwise margins but lacks pooled decline. Their partial successes are a positive diagnostic result: the same-round exposure signature is compatible with accumulated social updating and cannot uniquely identify an age-like mechanism (Appendix E).

Cross-family scope. Yi-1.5-6B-Chat reproduces the exposure partial correlation (-0.061 $[-0.087, -0.031]$) but declines more slowly than Qwen. A frozen 64-population Mistral test passes its same-round (-0.00426 [-0.00755,-0.00090]) and null-placebo (-0.00036 [-0.00358,+0.00283]) gates. At $T = 0.7$, its early–late decline remains $+0.1509$ $[+0.1269, +0.1758]$, but the local

slope -0.00256 $[-0.00583, +0.00070]$ crosses zero. In the opposite-objective role task, Mistral recovers -0.00537 $[-0.00801, -0.00277]$; Qwen recovers a long-horizon contrast of $+0.5021$ $[+0.4792, +0.5245]$, while its local slope -0.00309 $[-0.00674, +0.00053]$ crosses zero; both placebos are null (Appendix O). Long-horizon decline spans three families, stochastic decoding, and two objectives, while local rate is not invariant.

Prospective window breakpoint. Finite memory gives a falsifiable change point: history grows before $e = H$ and turns over after it. Matched Qwen $H \in \{10, 30, 60\}$ role runs reveal a threshold contrast of $+0.3911$ $[+0.3561, +0.4270]$; this was discovery evidence. The same bins, schedules, contrast, and gate were frozen before Mistral outcomes: its contrast is $+0.1352$ $[+0.1064, +0.1632]$ and pooled rebound $+0.0618$ $[+0.0477, +0.0750]$, both positive with intervals excluding zero; the post-fill slope and placebo also pass. A 64-population, independent-seed Qwen arm prospectively repeats the rebound ($+0.2863$ $[+0.2641, +0.3078]$), while an archived naming panel repeats its direction (Appendix P). Thus the organizing pattern is not monotone age but whether finite evidence is accumulating or turning over.

Replay reference. An exposure-linked plasticity decline is not by itself a pathology: evidence accumulation can produce the same pattern. We therefore replay a transparent majority-count reference observer through *the same logged partner sequence each real agent experienced*, decision by decision. The observer tracks partner counts (c_1, c_2) and chooses the majority name (ties keep the previous choice). Under a stationary Bernoulli model with any symmetric Beta prior, deterministic posterior-predictive argmax reduces to majority counting. This makes the benchmark parameter-free, not optimal for the non-stationary environment. It also counts the agent's entire logged sequence whereas the LLM sees only $H = 30$ events, and replay lacks feedback because its stream was generated by the real population (Limitations).

Table 4: $P(\text{switch} \mid \text{last partner disagreed})$ by own exposure: the LLM against the replayed reference observer, both regimes.

exposure e	1–4	4–7	7–10	10–14	14–18	18–24	24–32
free convergence							
reference observer	0.522	0.137	0.202	0.089	0.073	0.060	0.045
LLM	0.327	0.107	0.092	0.062	0.028	0.011	0.007
LLM / reference	0.63	0.78	0.46	0.70	0.38	0.19	**0.15**
12.5% committed minority							
reference observer	0.540	0.143	0.192	0.112	0.063	0.039	0.038
LLM	0.357	0.113	0.113	0.109	0.034	0.018	0.010
LLM / reference	0.66	0.79	0.59	0.98	0.54	0.45	**0.25**

The reference observer's plasticity falls from 0.52 to 0.045, showing that accumulated evidence alone produces a large decline. The LLM switches at ~ 0.65 of the reference rate when young and 0.15–0.25 beyond $e \approx 24$; its replay-relative rigidity ratio is therefore 4–$6\times$ at high exposure. This asymmetric reference contextualizes the magnitude, while Section 4.4 independently holds the incoming challenge fixed.

Candidate history summaries. Conversational inertia (Wan et al., 2026) and action-history bias (Zhang et al., 2026) motivate asking whether agents continue their own past choices. At fixed early exposure, plasticity falls from 0.35 for own-choice streaks of 1–2 to 0.11 for streaks of 3–5; mixed-evidence states also remain rigid relative to the replay reference. In a forced-deviation arm, following a partner away from the agent's long-run majority rises to 0.156 $[0.112, 0.200]$ from an exposure-matched baseline of 0.009 $[0.005, 0.014]$. The intervention moves choices, rewards, feedback, and later population state together, making history-mediated closed-loop persistence the supported system-level interpretation and streak length one candidate summary.

History interventions and reset probes. On a fixed panel of 618 Stage 6 snapshots, reward-consistent flipping of the earlier partner prefix increases alternate-name score by $+0.037$ ($[+0.009, +0.067]$) in the primary cohort, and an off-distribution own-choice flip produces $+0.193$

$[+0.137, +0.254]$. Across Stages 7B–7D, window and temperature changes, one-time truncation, and same-state dual-window scoring do not meet their frozen practical-recovery gates; detailed arm estimates are reported in Appendices F–I. Together, the Stage 7 panel establishes decision sensitivity to history content while defining the current control frontier: the tested reset probes do not yet provide a reliable recovery policy.

4.4 Act IV—Functional diagnosis under a common challenge

The first three acts establish why neither shared round nor exposure alone finishes the attribution. We therefore take $H = 30$ snapshots from the verified index-free trajectories at $e \in \{32, 64\}$ and replay each twice for $K = 40$ events: an exogenous partner always emits the incumbent name (support) or the alternative (challenge). The same four agents in each of 24 retained populations contribute both exposures. Model actions are real, rewards are recomputed, and histories update normally. Response area U is the challenge-minus-support alternative-choice mean over steps 1–40; the pre-specified comparison is $U_{64} - U_{32}$ with 5,000 paired-population bootstrap draws.

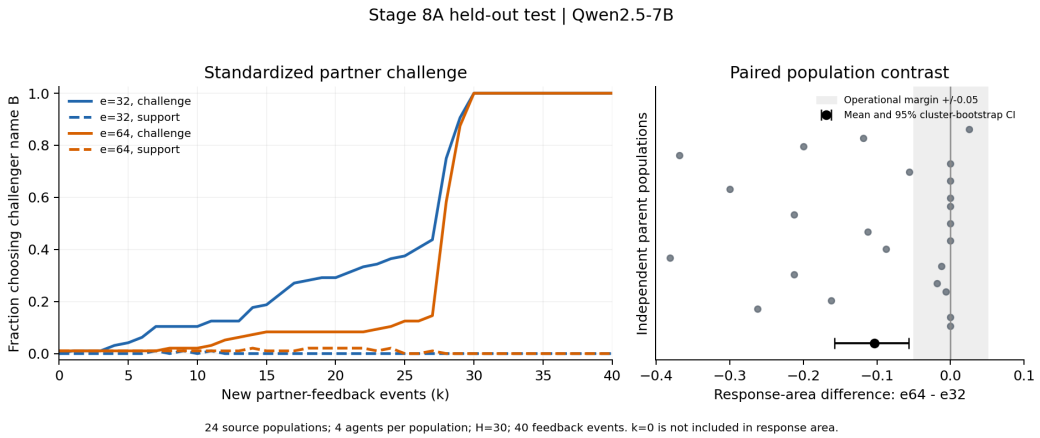

Figure 3: Standardized sustained challenge. Left: challenged-name choice fractions; support branches remain near zero and both challenge curves reach one by $k = 30$. Right: the pre-specified raw population-paired $U_{64} - U_{32}$ contrast. The grey band is a reporting margin, not a capability threshold.

The raw response area is 0.4471 at $e = 32$ and 0.3435 at $e = 64$, a difference of -0.1036 $[-0.1578, -0.0565]$. All 96 challenged snapshots in each group nevertheless choose the challenged name by $k = 40$, both groups have median sustained-adoption onset 28, and both challenge curves reach one by $k = 30$. The raw contrast therefore concerns response speed, while eventual adaptation is preserved.

The histories also contain different evidence: visible partner support for the incumbent averages 0.744 versus 0.910. After observing this imbalance, we ran the four-rung post-hoc common-support analysis in Table 5. It is a sensitivity analysis, not a causal or equivalence test: it maps attenuation under measured evidence balance and leaves history order for prospective study. Every measured-history specification yields a near-zero within-stratum contrast.

Table 5: Stage 8B common-support sensitivity. Contrasts are $U_{64} - U_{32}$; e32/e64 gives retained snapshots.

exact history summaries	contrast	95% interval	e32/e64
partner evidence count	$+0.0003$	$[-0.0117, +0.0113]$	70/87
+ own count, success count	$+0.0002$	$[-0.0117, +0.0152]$	58/75
+ last-event indicators	$+0.0076$	$[-0.0036, +0.0230]$	52/73
+ exact transition counts	-0.0027	$[-0.0082, +0.0054]$	35/65

The attenuation is not a response floor. The first rung retains 70/96 $e = 32$ snapshots (mean $U = 0.3807$) and 87/96 $e = 64$ snapshots (mean $U = 0.3434$); set-aside early histories have higher response and weaker incumbent evidence. Thus the raw gap concentrates in early histories without later-history counterparts and in between-stratum evidence composition. We next test the remaining within-window coordinate prospectively. After an inconclusive early endpoint, a frozen Qwen replication holds all 30 event tuples and the final three positions fixed and finds that later counter-evidence increases $k = 1$–40 response area by +0.1723 [+0.1254,+0.2203]. A matched 64-population Mistral test yields -0.0121 [-0.0344,+0.0092] and fails the same directional gate (Appendix K). Order is therefore a functional Qwen coordinate here, not a model-invariant direction.

4.5 Act V—Source localization and transmission bottleneck

We reconstruct full 12-agent checkpoints from held-out $e = 64$ trajectories and create paired branches differing only in one source history. The edit swaps both names in 27 of 30 visible events while preserving the last three, exposure, order, and success/failure sequence; all other histories and future schedules are identical (24 populations, two sources each).

The source next-action effect is 0.708 $[0.542, 0.854]$. After one contact, the difference reaches 0.699 $[0.597, 0.784]$ of non-source memories, but only 1/48 cases ever differs in a non-source action and no late state difference remains. With six contacts per chain edge, 4/48 cases diverge downstream, all after the controlled prefix; relative to one contact, non-source action-divergence rate rises by $+0.00231$ $[+0.00010, +0.00498]$ and late state Hamming by $+0.00398$ $[+0.00028, +0.00900]$. This is a strong source-history effect followed by a high-attenuation, dose-sensitive transmission bottleneck, not an immediate hop-by-hop cascade (full assays and the non-monotone post-hoc 12-contact extension: Appendix L).

A frozen Mistral replication (64 populations; 128 cases) misses the immediate source-localization gate ($+0.094$ $[-0.008, +0.195]$) but yields any non-source divergence in 0.844 $[0.781, 0.898]$ of cases at a sparse 0.01766 $[0.01532, 0.02012]$ action rate; population-share effects remain null. Qwen's strong-source/high-attenuation profile is therefore not a universal propagation law (Appendix M).

5 Discussion and implications

From symptom to transmission. The five acts separate shared time, histories, evidence/order, standardized response, and propagation. DriftClock diagnoses states, not traits: challenges preserve updating, while paired edits expose model-dependent source sensitivity, action leakage, and population displacement—diagnostic outputs rather than one presumed universal law.

For diagnosis. Report t, e, and z; use exact-time separation, history normalization/reordering, and common challenges. Treat $e = H$ as a prospective breakpoint: matched window arms distinguish cumulative exposure from evidence turnover. Then branch one source and inspect downstream action boundaries before replacement.

6 Conclusion

DriftClock turns a population-ageing curve into an auditable localization sequence. Shared time reveals the symptom; exposure localizes heterogeneous histories; matched window interventions expose a cross-model filling-to-turnover breakpoint; evidence composition and order characterize functional differences; paired branches test whether a localized source propagates. Diagnosing what changed, where it resides, how finite history maintains it, and whether it spreads requires this sequence.

7 Limitations

Frozen weights, prompt-only state, three near-7B families, mostly argmax decoding, two binary tasks, and one asynchronous scaffold constrain causal interpretation and universality (Appendices E–P).

ETHICS STATEMENT

Simulated populations of open-weight models; no human subjects. Measuring how response varies across interaction histories can inform monitoring as well as manipulation; we release nothing beyond standard simulation tooling.

REPRODUCIBILITY STATEMENT

All LLM experiments are inference-only on open-weight models. We release simulators, configurations, complete decision logs, analyses, seeds, manifests, exit markers, and SHA-256 provenance for the clock, placebo, collapse, social-baseline, and Stage 7 tests. Stage 8A adds its freeze manifest, protocol, 192 snapshots, and complete 15,744-decision log (SHA-256 `467451948f41148bbb888a7ef3a45406d48f0825deecec6bc34215b10a7a3842`); its primary was frozen before retained-population responses were queried. Stage 8B is a no-inference post-hoc sensitivity analysis. Stage 9 releases complete 23,040-decision logs for one- and six-contact assays, frozen packages, source checkpoints, and pre-output non-source audits; the 12-contact extension is post-hoc. The order study releases the initial 17,384-decision assay, its failed early primary, and the separately frozen 10,496-decision Qwen replication. Stages 10B–C add 51,200-decision Mistral clock trajectories and a frozen 20,992-decision order assay over all 64 populations. Stage 10D adds 61,440 paired-branch decisions; Stage 10E adds 51,200 decisions under frozen $T = 0.7$ sampling. Stages 10F–G add 102,400 complementary-role decisions across Mistral and Qwen. Stage 10H adds a 128,000-decision Qwen matched-window panel; Stages 10I–J add the independently frozen 128,000-decision Mistral panel; Stage 10K adds a 76,800-decision independent-seed Qwen replication. Formal intervals use 5,000 population-bootstrap draws. Exact same-round, social-baseline, and initial Qwen turnover analyses are review-motivated; the Mistral threshold and Qwen replication tests were frozen before their outcomes. Local protocols are not public preregistrations.

AI USE DISCLOSURE

Generative AI tools (OpenAI Codex) assisted literature discovery, experimental-design feedback and robustness-check formulation, code development, data and statistical auditing, result interpretation, and language editing. The authors reviewed all AI-assisted work, executed or supervised the experiments, tested AI-assisted code, checked every reported result against the underlying logs and every cited claim against its source, and retain full responsibility for the submission.

REFERENCES

A. Al-Tawaha, S. Gu, P. Niu, R. Jia, and M. Jin. Remembering more, risking more: Longitudinal safety risks in memory-equipped LLM agents. *arXiv:2605.17830*, 2026.

A. F. Ashery, L. M. Aiello, and A. Baronchelli. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20):eadu9368, 2025.

A. Baronchelli, M. Felici, V. Loreto, E. Caglioti, and L. Steels. Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics*, P06014, 2006.

C. Barrie and P. Törnberg. Emergent LLM behaviors are observationally equivalent to data leakage. *arXiv:2505.23796*, 2025.

A. Backlund and L. Petersson. Vending-Bench: A benchmark for long-term coherence of autonomous agents. *arXiv:2502.15840*, 2025.

A. C. Cameron and D. L. Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015.

D. Centola, J. Becker, D. Brackbill, and A. Baronchelli. Experimental evidence for tipping points in social convention. *Science*, 360(6393):1116–1119, 2018.

M. Cemri, M. Z. Pan, S. Yang, et al. Why do multi-agent LLM systems fail? In *NeurIPS Datasets and Benchmarks*, 2025. arXiv:2503.13657.

M. Chen, J. Wang, F. Mu, Y. Wang, Z. Liu, H. Feng, and Q. Wang. Seeing the whole elephant: A benchmark for failure attribution in LLM-based multi-agent systems. In *ACL*, pages 19888–19905, 2026.

G. De Marzo, A. Bellina, C. Castellano, V. Priesemann, and D. Garcia. Conformity generates collective misalignment in AI agents societies. *arXiv:2605.10721*, 2026.

D. Guilbeault, S. Caplan, and C. Yang. A simple threshold captures the social learning of conventions. *PNAS*, 2026. doi:10.1073/pnas.2508061123.

K. Li, T. Liu, N. Bashkansky, D. Bau, F. Viégas, H. Pfister, and M. Wattenberg. Measuring and controlling instruction (in)stability in language model dialogs. In *COLM*, 2024. arXiv:2402.10962.

P. Laban, H. Hayashi, Y. Zhou, and J. Neville. LLMs get lost in multi-turn conversation. In *ICLR*, 2026. arXiv:2505.06120.

A. Mehdizadeh and M. Hilbert. Exploring the topology and memory of consensus: how LLM agents agree, fragment, or settle when forming conventions. *arXiv:2606.04197*, 2026.

M. Okawa. Emergence of biased consensus in multi-agent LLM debates. In *ICML*, 2026. arXiv:2608.02827.

H. Tanaka. When is collective intelligence a lottery? Multi-agent scaling laws for memetic drift in LLMs. *arXiv:2603.24676*, 2026.

D. Yang. When is emergent consensus real? A measured coupling gain and a validity diagnostic for LLM agent societies. *arXiv:2606.22203*, 2026.

Y. Wan, Z. Cao, Z. Zhang, Z. Zeng, S. Shen, C. Meng, and L. Zhu. Mitigating conversational inertia in multi-turn agents. In *ICML*, 2026. arXiv:2602.03664.

S. Zhang, M. Yin, J. Zhang, et al. Which agent causes task failures and when? On automated failure attribution of LLM multi-agent systems. In *ICML*, pages 76583–76599, 2025.

X. Zhang, Z. Zhang, Z. Fu, B. Lin, and W. Wang. From inertia to objectivity: Improving deep research agents with noise isolation. *arXiv:2608.23045*, 2026.

A PLASTICITY DEFINITIONS

Three definitions, identical shape, different levels. Raw switch rate: $0.118 \rightarrow 0.000$ (free) and $0.132 \rightarrow 0.000$ (committed) across the exposure bins of Table 3. Agent-observable conditioning (main text): $0.271 \rightarrow 0.000$ and $0.296 \rightarrow 0.000$. Collider version, $P(\text{switch} \mid \text{THIS round's partner disagreed})$: $0.154 \rightarrow 0.000$ and $0.184 \rightarrow 0.000$—reported only to show the collider does not drive the result; it conditions on the same round's outcome and is not used for inference.

B ESTIMATOR DETAILS

The main exposure estimate uses exact (population run $\times$ round) cells. Each cell contains the two agents selected at that round; differencing fixes calendar round and pre-decision population state. The slope is $\sum_c \Delta e_c \Delta y_c / \sum_c (\Delta e_c)^2$ over unequal-exposure pairs. For previous-disagreement switching, widening cells from 1 to 5, 10, and 25 rounds gives slopes $-0.00659/-0.00500/-0.00545/-0.00575$ in free convergence and $-0.00550/-0.00810/-0.00834/-0.00915$ under the committed minority; every 95% interval excludes zero. Joint run-by-25-round-epoch models give population-state support coefficients -0.02222 $[-0.03202, -0.01333]$ and -0.04006 $[-0.05286, -0.03011]$. The collective-clock specification holds run and integer exposure exactly fixed; its residual round and state estimates appear in §4.1. All intervals are percentile cluster bootstraps resampling whole population runs (600 draws; 32 runs per regime). The 768/672-cell within-agent contrast in §4.2 uses *unconditional* switching in each agent's first versus fourth exposure quartile, requiring at least three decisions on each side; it is not the conditioned plasticity statistic.

C RANDOMIZATION AND PLACEBO AUDIT

For the randomization test, pair-level numerators $\Delta e_c \Delta y_c$ are aggregated within each of the 32 population runs. We independently flip the sign of each run aggregate in 20,000 Monte Carlo draws and recompute the slope using the observed denominator. The two-sided p-value is the share of null slopes at least as extreme as the observed slope, with the standard plus-one correction. For the negative control, future exposure is the number of subsequent interactions an agent receives between the current decision and the end of its run. Replacing accumulated exposure with this future count preserves heterogeneity in eventual activity but reverses temporal order: future pair selections cannot cause the current decision. Current-exposure plasticity slopes are -0.00659 ($p < 10^{-4}$) and -0.00550 ($p = 0.0033$); future-exposure slopes are $+0.00036$ ($p = 0.81$) and $+0.00055$ ($p = 0.75$). For unconditional switching, future-exposure slopes are $+0.00034$ ($p = 0.48$) and $+0.00050$ ($p = 0.34$).

D SECONDARY ABLATION AND INTERVENTION DETAILS

Renderer ablation. The baseline renderer prints each agent's private interaction indices ("round n:"), making its clock legible. We re-ran free convergence without them (32 populations $\times$ 400 rounds). Plasticity still falls $\sim$50$\times$ along exposure ($0.246 \rightarrow 0.005$), and the within-agent contrast is -0.047 (240 cells down versus 26 up). The separately estimated within-epoch gradient is -0.049 $[-0.059, -0.040]$ with indices and -0.017 $[-0.034, +0.005]$ without; a population-clock coefficient is -0.020 $[-0.036, -0.007]$ index-free. Because the renderer contrast itself was not paired, these estimates do not isolate an amplification effect. Stage 6 strengthens index-free robustness without isolating count from content.

Mid-run perturbation. We injected 6 of 24 agents, committed to the name opposite each population's current majority, at round 50 or 300 in free-running populations (16 per arm). The early arm runs 280 rounds and the late arm 530, giving equal 230-round post-injection windows; these horizons differ from the 400-round core runs. At matched time since injection (the first 100 post-injection rounds), free agents follow committed pushers at 0.041 $[0.024, 0.058]$ early versus 0.003 $[0.000, 0.007]$ late. At matched exposure (20–30), the rates are $0/340$ versus 0.005 $[0.000, 0.012]$. Neither arm produces takeover ($0/16$ each), and free-agent adoption remains flat over 230 rounds

($0.257 \rightarrow 0.253$ early; $0.149 \rightarrow 0.115$ late). These endpoints do not establish an effective correction window.

E INDEX-FREE CONDITIONAL AND SOCIAL-BASELINE AUDIT

The following analyses were added after the frozen collapse result. They use the same exact same-round estimator with 5,000 population-cluster bootstrap draws. The fresh $N{=}24$ result is a post-hoc exact-revision sensitivity; the archived arm remains part of the frozen gate.

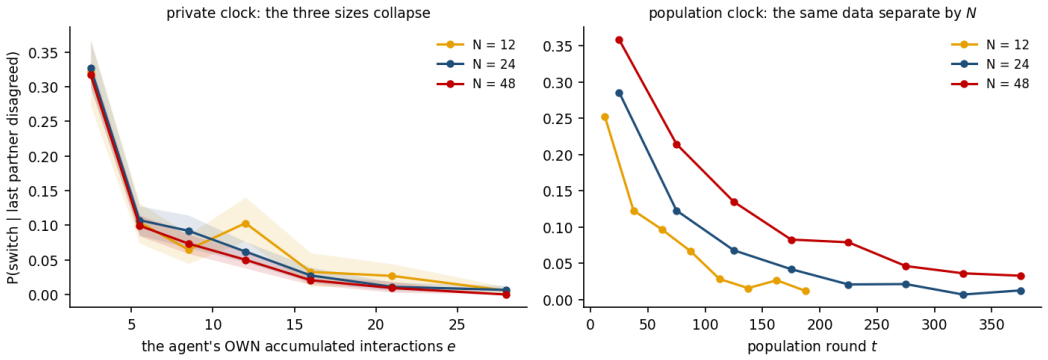

Figure 4: Standard-renderer scale diagnostic (32 populations per size; 200/400/400 rounds). Exposure-indexed curves approximately align, while the identical decisions separate on the round axis. This exploratory diagnostic is distinct from the pre-specified index-free 200/400/800-round test in Figure 2.

Table 6: Index-free exact same-round previous-disagreement slopes.

arm	slope	95% interval	unequal-exposure pairs
archived $N{=}24$	-0.0007	[-0.0026, 0.0018]	2,109
exact-revision $N{=}12$	-0.0027	[-0.0057, -0.0001]	781
exact-revision $N{=}24$ (post-hoc)	-0.0023	[-0.0038, -0.0009]	2,123
exact-revision $N{=}48$	-0.0030	[-0.0043, -0.0018]	4,108

Table 7: Exact same-round slopes for social-history baselines. Asterisks mark 95% intervals excluding zero. None passes the complete collapse gate.

rule	$N{=}12$	$N{=}24$	$N{=}48$	full gate
sliding majority	-0.0513^{*}	-0.0192^{*}	-0.0194^{*}	no: $N{=}12$ late cell empty
inertial majority	-0.0082	-0.0021	-0.0025^{*}	no: pooled decline null
EWMA social state	-0.0630^{*}	-0.0399^{*}	-0.0189^{*}	no: $N{=}12$ late cells empty

F STAGE 7 PAIRED HISTORY INTERVENTION

The Stage 7A panel was scored on Qwen2.5-7B-Instruct at revision `a09a35458c702b33eeacc393d103063234e8bc28`. It contains 618 snapshots (161 from $N = 12$, 204 from fresh $N = 24$, and 253 from $N = 48$); all 32 primary populations contribute. Three of the 64 primary run-by-anchor cells had no eligible snapshot, and this coverage limitation is retained. Each snapshot supplies its original 30-event history, a last-10 truncation, a shuffle of the earlier 27 events, a partner-choice flip with rewards recomputed, and an own-choice flip with rewards recomputed; the last three events remain fixed. Both option orders are evaluated. Thus the prefix flips are joint history-and-reward interventions, not isolated streak manipulations.

Table 8: Post-hoc binwise disclosure for the exact-revision Stage 6 sensitivity. The pre-specified gate uses count-weighted mean absolute deviation, not the maximum single-bin difference.

exposure bin	N=12	N=24	N=48	maximum absolute difference
$[1, 4)$	0.2673	0.2518	0.2862	0.0344
$[4, 7)$	0.0537	0.0358	0.0322	0.0215
$[7, 10)$	0.0376	0.0151	0.0205	0.0226
$[10, 14)$	0.0167	0.0094	0.0094	0.0074
$[14, 18)$	0.0073	0.0116	0.0108	0.0042
$[18, 24)$	0.0066	0.0023	0.0066	0.0044

The primary outcome is the alternate-name probability from a two-option $T = 0.7$ softmax over the two name logits. The LLM estimates below average option order and snapshot within population and use 5,000 population-cluster bootstrap draws. The familywise intervals are 98.333% percentile intervals for the three primary contrasts; ordinary 95% intervals and deterministic argmax outcomes are retained in the machine-readable analysis artifact.

Table 9: Stage 7A LLM contrasts versus the original prompt. Fresh $N = 24$ is primary; $N = 12$ and $N = 48$ are descriptive. Values are estimates and familywise intervals.

source	contrast	estimate	familywise CI
$N = 12$	trim to last 10	$+0.0103$	$[-0.0033, +0.0276]$
$N = 12$	shuffle earlier prefix	$+0.0104$	$[-0.0045, +0.0383]$
$N = 12$	flip partner prefix+reward	$+0.0737$	$[+0.0450, +0.1026]$
fresh $N = 24$	trim to last 10	$+0.0091$	$[-0.0084, +0.0305]$
fresh $N = 24$	shuffle earlier prefix	$+0.0190$	$[-0.0030, +0.0428]$
fresh $N = 24$	flip partner prefix+reward	$+0.0370$	$[+0.0095, +0.0673]$
$N = 48$	trim to last 10	$+0.0116$	$[-0.0009, +0.0277]$
$N = 48$	shuffle earlier prefix	$+0.0130$	$[+0.0004, +0.0274]$
$N = 48$	flip partner prefix+reward	$+0.0404$	$[+0.0221, +0.0617]$

All three frozen reference policies use a deterministic switch indicator rather than a calibrated probability, so their effect sizes are not directly comparable with the LLM. On the same primary snapshots, their partner-prefix contrasts are:

Table 10: CPU reference-policy response to the partner-prefix/reward flip in fresh $N = 24$. The outcome is a deterministic switch indicator, not an LLM probability.

policy	estimate	familywise CI
sliding majority	$+0.3788$	$[+0.1768, +0.5684]$
inertial majority	$+0.3554$	$[+0.1565, +0.5464]$
window-reset EWMA	$+0.2482$	$[+0.1084, +0.4035]$

The common positive direction is evidence that partner-history sensitivity is not unique to the LLM; it is not evidence that these three rules reproduce the LLM's complete dynamics. The LLM's own-prefix contrast is secondary ($+0.1934$, familywise interval $[+0.1370, +0.2543]$) and remains a hypothesis-generating off-distribution response.

The full result, manifest, source panel, reference-policy logs and analyses, frozen protocol, and execution code are released together. The scored LLM JSONL has SHA-256 `4afe4e180c8360011298167b816d40a42346adfb03401498b3dd34df47421425`. Stage 7B closed-loop window and temperature arms are reported separately; Stage 7A must not be read as evidence of recovery, takeover resistance, or a unique private clock.

G STAGE 7B CLOSED-LOOP WINDOW AND TEMPERATURE TEST

The prospective local protocol compares four closed-loop Qwen2.5-7B-Instruct arms: H=30, T=0 (reference), H=10, T=0, H=60, T=0, and H=30, T=0.7. Each arm has 32 independently evolv-

ing populations of 12 agents for 720 rounds, with two decisions per population per round. The rotating-neighbourhood scheduler, option-order stream, and random-number stream are seeded identically across arms; choices and subsequent histories are allowed to diverge. Memory contains the last H interactions. The stochastic arm samples from the two-option softmax using the frozen choice stream; deterministic arms use argmax. This is a closed-loop sensitivity test, not a randomized treatment of memory at fixed realized social history.

The primary endpoint is each population's unconditional own-choice switch rate among decisions with prior exposure $60 \leq e < 120$. We first average decisions within population and then take paired differences from $H{=}30, T{=}0$ for the three contrasts. A population contributes only with at least 40 decisions in the endpoint window; at least 24 paired populations are required for a supported primary contrast. Uncertainty uses 5,000 population-paired bootstrap draws, with familywise 95% coverage across the three contrasts. The frozen analysis additionally reports raw-exposure, exposure-to-window, and visible-history curves; conditional switching after a partner disagreement; and exact same-round exposure slopes. Sparse bins are labelled inconclusive rather than treated as zero. The full run has 46,080 decisions per arm. Seeds, thresholds, bins, and inclusion rules are fixed in `protocol.json`; this local protocol was not a public preregistration.

Table 11: Stage 7B frozen primary endpoint. Differences are arm minus $H{=}30, T{=}0$; intervals are familywise 95% population-paired bootstrap intervals over the three planned contrasts.

contrast	paired populations	difference	familywise interval
$H{=}10, T{=}0$	32	-0.001246	$[-0.003349, +0.000224]$
$H{=}60, T{=}0$	32	$+0.002597$	$[+0.000170, +0.005317]$
$H{=}30, T{=}0.7$	32	$+0.000900$	$[-0.000725, +0.003014]$

Shortening the window to $H{=}10$ changed the late-exposure switch rate by -0.001246, with a familywise interval that crosses zero; this is no detected difference and does not establish equivalence. Enlarging the window to $H{=}60$ increased the endpoint from 0.001573 to 0.004170, a difference of $+0.002597$ ($[+0.000170, +0.005317]$), or approximately 0.260 percentage points. The absolute effect is small, but its familywise interval excludes zero. At $H{=}30, T{=}0.7$, the endpoint was 0.002473, a $+0.000900$ difference from deterministic decoding ($[-0.000725, +0.003014]$); this is also no detected difference, not equivalence.

These are closed-loop arm contrasts. Scheduler and option-order streams are matched, but treatment-dependent choices alter rewards, partner feedback, population state, and later histories. The estimates therefore quantify whole-system sensitivity to a window or decoder configuration; they do not isolate memory length or temperature while holding realized social history fixed. The $H{=}10$ result is not an equivalence claim, and the $H{=}60$ result does not establish a monotone window-length law.

The descriptive exact same-round slopes are -0.002674 ($[-0.006173, -0.000723]$; 905 unequal-exposure pairs) for $H{=}10, T{=}0$, -0.001411 ($[-0.003436, -0.000092]$; 1,408 pairs) for $H{=}30, T{=}0$, -0.000128 ($[-0.000839, +0.000539]$; 2,355 pairs) for $H{=}60, T{=}0$, and -0.001884 ($[-0.004076, +0.000148]$; 1,311 pairs) for $H{=}30, T{=}0.7$. No between-arm slope contrast was frozen, so the apparent attenuation is descriptive only. Raw-exposure, exposure-to-window, and visible-history curves likewise do not select a unique alignment coordinate.

H STAGE 7C PAIRED VISIBLE-HISTORY TRUNCATION CHALLENGE

Frozen design. Stage 7C forks the verified `h30_t0` terminal checkpoints at source round 720. The frozen source contains 32 parent populations with 12 agents each; every included endpoint must satisfy the frozen unanimity and provenance checks. Within each parent, a fixed subset of 3 agents is committed to the challenger and the remaining 9 agents enter the paired branches: `reset0` retains their visible histories, `reset5` truncates the frozen nested subset, and `reset9` truncates all free-agent histories to the latest one complete interaction record. Truncation preserves cumulative exposure, current choice, and the retained interaction; the index-free recent-history renderer, mem-

ory cap, decoding rule, and post-fork horizon remain fixed at $H{=}30$, deterministic argmax ($T{=}0$), and 230 rounds.

For each parent, the branches share a frozen post-fork schedule and option-order stream. Different parents use independent streams, and branch execution order is prospectively rotated. R2 repeats the same endpoint, branches, estimand, inclusion rules, and gate using independently frozen committed-agent, reset-agent, schedule, branch-order, and bootstrap streams. It reuses the same parent checkpoints as R1 and therefore does not add independent populations. R2 was required to run regardless of the R1 result, and no result-selected additional randomization is permitted.

Primary estimand and gate. Let F_p be the frozen free-agent set for parent p, W the frozen primary post-fork round window, and c_p its challenger label. For branch b and randomization replicate q, the parent-level endpoint is

$$Y_{pbq} = \frac{\sum_{r \in W} \sum_{i \in F_p} \mathbf{I}\{C_{pbqir} = c_p\}}{|W|\,|F_p|}, \tag{2}$$

where C_{pbqir} is free agent i's latest choice after post-fork round r. The replicate-specific primary contrast is $\Delta_{pq} = Y_{p,\texttt{reset9},q} - Y_{p,\texttt{reset0},q}$ and its estimate averages Δ_{pq} over complete parent populations. Each interval uses 5,000 paired parent-population bootstrap draws. Formal analysis requires all 32 complete parent triples; any missing or incomplete triple makes that replicate inconclusive rather than creating a reduced-sample success analysis.

Writing L_q and U_q for the replicate-specific confidence limits and $\tau = 0.05$ for the frozen practical threshold, the frozen gates are: practical positive when $L_q > \tau$; detectable but below the established practical magnitude when L_q is positive but no greater than τ; negative when U_q is negative; and inconclusive otherwise, including any interval spanning zero. An inconclusive result is not equivalence. The `reset5` branch is a descriptive saturation check, not a linear agent-dose or proportionality test; all late-window, point-share, adoption, decision-level, and majority-crossing analyses are secondary.

R1/R2 combination rule. For the frozen replicate set $\mathcal{R} = \{\mathrm{R1}, \mathrm{R2}\}$, the combined parent contrast is

$$\bar{\Delta}_p = \operatorname*{mean}_{q \in \mathcal{R}} \Delta_{pq}. \tag{3}$$

The combined analysis bootstraps parent IDs while carrying both randomizations together; it never pools R1 and R2 as independent clusters. It reports each replicate, the mean of $\bar{\Delta}_p$, and the within-parent seed-sensitivity contrast $\Delta_{p,\mathrm{R2}} - \Delta_{p,\mathrm{R1}}$. The combined interval is a conditional precision summary over the frozen R1/R2 randomizations of the same parent checkpoints and cannot rescue a failed replicate. A practical effect may be described as independently reproduced only if both seed-specific lower confidence bounds exceed τ.

Table 12: Stage 7C frozen primary and combined analyses. Quantitative result fields are populated only from verified artifacts.

analysis	estimate	confidence interval	gate	role
R1	$+0.001667$	$[0.000000, +0.005000]$	inconclusive	seed-specific
R2	0.000000	$[0.000000, 0.000000]$	inconclusive	seed-specific
within-parent mean	$+0.000833$	$[0.000000, +0.002500]$	inconclusive	precision summary
R2 minus R1	-0.001667	$[-0.005000, 0.000000]$	—	sensitivity

The strict R1/R2 reproduction flag is `false`. Across the two prospectively frozen randomizations, one-time truncation did not establish restored population-level susceptibility under this challenge: R1 was driven by sparse adoption in one of 32 parents, while every R2 parent contrast was zero. Both seed-specific intervals fail the 0.05 practical threshold, and the combined interval also touches zero. This is an inconclusive result under the frozen gate, not an equivalence finding; it directly constrains any claim that memory clearing is a validated control. R1 and R2 use exit-zero, 44,160-row formal HF artifacts with decision-data SHA-256 values beginning `6f279c13` and `71b0c369`, respectively; the full digests, protocols, manifests, failed pre-inference R2 attempt, and combined analysis (SHA-256 beginning `f963785d`) are released.

Causal boundary. The intervention is a fork-time truncation of visible interaction history under a fixed committed-challenger pressure. Its contrast is the total closed-loop effect of that initial intervention: treatment-dependent actions change rewards, partner feedback, population state, and all later histories. It is therefore not a controlled direct effect of memory length or content at later rounds, not an intervention on cumulative exposure or an invisible age variable, and not a private-clock reset. Interference among free agents is part of the population-level treatment; agents, decisions, rounds, and the R1/R2 randomizations are not independent inferential units. Generalization is conditional on the frozen model, renderer, task, unanimous terminal states, challenger fraction, and post-fork horizon.

I STAGE 7D SAME-STATE DUAL-WINDOW TEST

Frozen design. Stage 7D tests whether the small Stage 7B $H{=}60$ closed-loop contrast can be accompanied by an immediate prompt-window effect when the realized pre-decision state is held fixed. It starts from the same 32 unanimous `h30_t0` terminal populations and reuses R2's frozen committed-agent assignment and the first 100 rounds of its schedule. Three agents per population are committed to the challenger; the other nine remain free. At each of the 4,847 free-agent decision states, the model scores the identical pre-decision lifetime history under $H{=}30$ and $H{=}60$, each in both option orders. All four scores are computed before either member of the scheduled pair is updated.

A prospectively frozen bit assigns either $H{=}30$ or $H{=}60$ as the executed window for each free decision. Only that window's scheduled-normal-order deterministic argmax enters the environment; the other three scores are counterfactual measurements. Thus each parent follows one mixed shared trajectory rather than two diverging pure-window trajectories. The assignment contains 3,164 $H{=}30$ and 3,236 $H{=}60$ slots overall (2,403 and 2,444, respectively, among free decisions) and was frozen before formal model outputs. The experiment comprises 19,388 prompt evaluations and 6,400 logged pair-side rows.

Estimand and inference. For each free decision state, the primary contrast is the mean over the two option orders of the deterministic $0/1$ challenger indicator under $H{=}60$ minus the corresponding mean under $H{=}30$. We average these state contrasts within each parent and then equally across the 32 parents. A 95% interval uses 5,000 parent-population bootstrap draws; any missing parent invalidates formal analysis. The frozen practical threshold is $\tau = 0.01$: practical positive requires a lower bound above $+0.01$, detectable but sub-threshold positive has a lower bound in $(0, 0.01]$, negative requires an upper bound below zero, and all other cases are inconclusive. The secondary endpoint replaces the deterministic indicator with the two-option token-logit-renormalized challenger probability at $T{=}0.7$ and has no success gate.

Table 13: Stage 7D same-state $H{=}60 - H{=}30$ contrasts. The probability endpoint is secondary.

endpoint	parents	estimate	95% interval	gate
deterministic challenger choice	32	$+0.000214$	$[0.000000, +0.000642]$	inconclusive
$T{=}0.7$ challenger probability	32	$+0.000254$	$[+0.000025, +0.000671]$	descriptive

The primary interval touches zero and lies far below the frozen 0.01 practical threshold, so Stage 7D does not establish an immediate deterministic window effect on the same pre-decision states. The secondary probability contrast is positive but approximately 0.00025 and had no success gate. In combination with Stage 7B's small positive closed-loop switch-rate contrast, these results do not support reading the total trajectory difference as a demonstrated direct prompt-window effect; treatment-dependent feedback-path divergence remains a plausible contributor.

Causal boundary. This is a controlled immediate prompt-window contrast on the mixed shared-trajectory state distribution. It is not the total effect of deploying a pure $H{=}30$ or $H{=}60$ policy, and its challenger-choice outcome is not a numerical decomposition of Stage 7B's late-exposure switch-rate contrast. A positive result would establish an immediate window-sensitive component in this setting, not a private clock; a negative result would not make history irrelevant; and an interval

spanning zero would be inconclusive rather than an equivalence finding. Prompt renderings, option orders, decisions, agents, and rounds are repeated measurements, not independent inferential units.

The formal data SHA-256 begins `fd7f2ed4` and ends `b16533d`; the complete digest, manifest, analysis, protocol, freeze manifest, execution code, queue log, and exit marker are released with the data.

J STAGE 8A STANDARDIZED CHALLENGE

The formal Stage 8A panel contains 24 retained parent populations, four fixed agents per parent, two source exposures, two branches, and 40 post-snapshot feedback events (15,744 scored decisions including snapshot records). The $e = 32$ and $e = 64$ snapshots use full $H = 30$ visible windows. For each snapshot, support always emits the incumbent name and challenge always emits the alternative; model choices and rewards then update the prompt. The response area U is the mean challenged-name indicator in challenge minus support, averaged over steps, agents, and then parent populations. The primary contrast and bootstrap seed were frozen before any retained-parent response was queried. A preceding eight-population calibration selected $K = 40$ after $K = 12$ showed a response floor and is excluded from the formal estimate.

Table 14: Stage 8A branch means and response area. Branch columns are components of U, not separate response measures. "Ever B" counts snapshots that choose the challenged name at least once.

source	challenge	support	U	B at $k = 40$	ever B in support
$e = 32$	0.4479	0.0008	0.4471	96/96	1/96
$e = 64$	0.3518	0.0083	0.3435	96/96	2/96
$e64 - e32$	-0.0961	$+0.0076$	-0.1036	–	–

The paired-population 95% interval for $U_{64} - U_{32}$ is $[-0.1578, -0.0565]$. Both challenge curves reach one by $k = 30$ and both groups have median onset 28 for the first three consecutive challenged-name choices. The discriminating window is therefore $k = 1 \ldots 29$: the final 11 steps entering U are structurally saturated. $K = 40$ was retained because the preceding $K = 12$ calibration had shown a response floor but could not reveal the later convergence. The visible partner-support share differs (0.744 versus 0.910), motivating Stage 8B.

Table 15: LLM minus transparent-reference response area on identical snapshots and exogenous streams. Level intervals are Bonferroni-adjusted over six contrasts; exposure-difference intervals are descriptive.

reference	$e = 32$	$e = 64$	exposure-difference contrast
sliding majority	-0.2820 $[-0.3237, -0.2313]$	-0.3109 $[-0.3232, -0.2943]$	-0.0289 $[-0.0669, +0.0055]$
inertial majority	-0.2513 $[-0.2938, -0.1995]$	-0.2854 $[-0.2971, -0.2693]$	-0.0341 $[-0.0721, +0.0003]$
window EWMA	-0.5052 $[-0.5713, -0.4307]$	-0.5906 $[-0.6190, -0.5419]$	-0.0854 $[-0.1331, -0.0422]$

All level contrasts exclude zero for these three stated rules. The exposure-difference contrast is imprecise for sliding and inertial majority and negative for EWMA. The result does not identify a content-independent LLM residual, a rationality deficit, or a universal slope. The formal LLM log has SHA-256 `467451948f41148bbb888a7ef3a45406d48f0825deecec6bc34215b10a7a3842`.

K STAGE 8B COMMON-SUPPORT SENSITIVITY

After observing the Stage 8A evidence imbalance, we used exact strata of the existing 192 snapshots to estimate a symmetric overlap-population contrast. Within each stratum, the e64-minus-e32 response-area difference is weighted by $2n_{32}n_{64}/(n_{32}+n_{64})$; 95% intervals resample the 24 source populations 5,000 times. All four specifications and their overlap are reported in Table 16; none was selected by its result.

Matching partner evidence alone leaves a success-count imbalance (26.7 versus 24.1 of 30 events), motivating the second rung's own-incumbent and success counts; agreement across all four rungs is therefore informative rather than redundant. Within each bootstrap replicate, common strata are re-determined after parent resampling: their number ranges from 3 to 14 (median 10). The intervals consequently combine sampling variation with variation in the retained subpopulation and are read as sensitivity summaries, not as precise intervals for one fixed overlap parameter.

Table 16: Post-hoc evidence-normalized Stage 8B sensitivity. "Pairs" is the maximum possible one-to-one overlap, not the estimand's sample size.

exact history summaries	contrast	95% interval	e32/e64	pairs
partner evidence count	$+0.0003$	$[-0.0117, +0.0113]$	70/87	54
+ own count, success count	$+0.0002$	$[-0.0117, +0.0152]$	58/75	42
+ last-event indicators	$+0.0076$	$[-0.0036, +0.0230]$	52/73	39
+ exact transition counts	-0.0027	$[-0.0082, +0.0054]$	35/65	32

As a paired audit, 29 same-agent e32/e64 pairs from 10 populations have equal partner-incumbent counts. Their difference is exactly zero in 27 of 29 pairs. Twenty-one have identical visible memories and byte-identical stored snapshot-prompt hashes; all 21 reproduce exactly equal U, validating the render–replay–scoring pipeline end to end. Six additional pairs have different rendered histories but unanimous 30-event partner evidence and also yield zero contrast. Only two pairs carry nonzero information ($+0.2500$ and $+0.0750$), each from a population contributing one pair; their equal-population mean is $+0.0325$ and pair-level mean is $+0.0112$. We therefore report this subset descriptively without an interval. The exact-stratum analysis maps attenuation under measured evidence balance while leaving order uncontrolled.

Prospective history-order intervention

For each eligible snapshot we stable-partition the first 27 visible event tuples by whether the partner supplied counter- or incumbent evidence, preserve within-type order, and leave the final three tuples fixed. Counter-early and counter-late arms therefore have identical exposure, exact event multiset, name counts, and success/failure counts. Each is branched again into incumbent support and alternate challenge for 40 feedback events under matched option order. In the initial 106-snapshot, 20-population assay, the frozen early $k = 1$–12 logit-margin difference-in-differences is $+0.1111$ $[-0.2477, +0.4657]$ and the corresponding binary estimate is +0.0112 [-0.0212,+0.0407]; both are inconclusive. The pre-labelled descriptive $k = 1$–40 binary estimate is $+0.1096$ $[+0.0690, +0.1493]$.

Before querying another model response, we froze that long-horizon binary endpoint for replication on 32 different Stage 6 source populations. Two exposure-32 snapshots per population are selected by an outcome-blind hash from all histories containing both evidence types (64 snapshots total). The population-bootstrap estimate is +0.1723 [+0.1254,+0.2203], passing the directional gate. Thus event order changes sustained challenge response beyond evidence counts in this assay. The original early primary remains a reported null; the replication supports a timescale-specific order effect rather than replacing it post hoc.

Prospective Mistral scope replication

Before formal population output, we fixed Mistral-7B-Instruct-v0.3 revision `c170c708c41dac9275d15a8fff4eca08d52bab71`, the calibrated single-token pair `coal`/`drum`, two 32-population cohorts, schedules, estimands, and gates. The first cohort reuses the Qwen scheduler seed but not its model output; the second uses an independent seed. Both fixed schedule hashes match, all 51,200 decisions parse, and the frozen pooled exact same-round slope is -0.00426 [-0.00755,-0.00090]. The early-minus-late contrast is +0.2713 [+0.2475,+0.2957], while the temporally reversed future-exposure placebo is -0.00036 [-0.00358,+0.00283]. The pooled primary negative-CI and placebo-contains-zero gates both pass. In a post-hoc cohort split, both slope estimates remain negative (-0.00462 and -0.00386); the first interval excludes zero, while the independent-seed cohort alone is less precise ($[-0.00888, +0.00140]$).

The Stage 10C order design was frozen while the first Stage 10B cohort was still running. It selects two exposure-32 snapshots from every population by outcome-blind hash (128 total), then applies the same exact-multiset, fixed-tail order contrast and 40-step paired challenge used for the Qwen replication. Its frozen binary endpoint is -0.0121 [-0.0344,+0.0092], so the directional lower-bound-above-zero gate fails. The early $k = 1$–12 contrast is -0.0371 [-0.0618,-0.0117], and the descriptive long-horizon logit contrast is -0.1088 [-0.2147,-0.0039]. Thus the shared exposure signature transfers to Mistral, while the Qwen ordering direction does not; the diagnostic profile is model-conditional rather than a universal order law.

L STAGE 9 PAIRED SOURCE AND PROPAGATION ASSAY

We reconstruct full 12-agent checkpoints from the verified Stage 7 $H = 30, T = 0$ log. A hash-ranked eight-population pilot is excluded from formal inference; the remaining 24 populations contribute two preselected source agents at exposure 64. Each case creates control and counter-history branches with identical initial histories except for the source edit, then uses identical future pairs and option-order bits. The edit swaps both own and partner names in the first 27 visible source events and preserves the last three, exposure, order, and the complete agreement/disagreement sequence. The intervention therefore changes visible name evidence without changing event-wise reward signs.

The one-contact protocol fixes source–direct, direct–two-hop, and two-hop–probe as the first three pairs, followed by a frozen uniform schedule. The separately frozen six-contact protocol repeats each chain edge six times before its frozen continuation. Both have 120 steps. The 12-contact arm was designed after seeing the six-contact output and is labelled exploratory. Formal intervals resample the 24 source populations 5,000 times after averaging their two cases.

Table 17: Stage 9 scheduled endpoints. Effects are treated-minus-control counter-name choices at the stated node; group quantities include the source.

contacts/edge	source at first contact	direct after source block	two-hop after direct block
1 (formal)	0.708 [0.542, 0.854]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
6 (formal)	0.688 [0.521, 0.854]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
12 (post-hoc)	0.708 [0.542, 0.854]	0.0208 [0.0000, 0.0625]	−0.0208 [−0.0625, 0.0000]

contacts/edge	group-share AUC	late group-share difference
1 (formal)	0.0620 [0.0487, 0.0738]	0.0549 [0.0391, 0.0693]
6 (formal)	0.0635 [0.0522, 0.0736]	0.0572 [0.0450, 0.0688]
12 (post-hoc)	0.0568 [0.0412, 0.0697]	0.0354 [0.0225, 0.0464]

Group-share contrasts in Table 17 mostly reflect persistence of the intervened source. We therefore froze a non-source audit during each formal run, before its output completed. It reconstructs both branches step by step and separately measures whether the other 11 agents have different visible memories, actions, or last-choice states.

Table 18: Non-source propagation audit. Action rate is the fraction of non-source decisions differing across paired branches; late metrics average the last 20 steps.

contacts/edge	any-action case fraction	action divergence rate	final memory reach
1 (formal)	0.021 [0.000, 0.063]	0.00019 [0.00000, 0.00057]	0.699 [0.597, 0.784]
6 (formal)	0.083 [0.021, 0.167]	0.00250 [0.00029, 0.00518]	0.646 [0.549, 0.733]
12 (post-hoc)	0.042 [0.000, 0.104]	0.00524 [0.00000, 0.01345]	0.585 [0.483, 0.672]

contacts/edge	late state Hamming	late counter-share difference
1 (formal)	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
6 (formal)	0.00398 [0.00028, 0.00909]	0.00398 [0.00038, 0.00900]
12 (post-hoc)	0.00379 [0.00000, 0.00947]	0.00000 [−0.00568, 0.00568]

Only one of 48 one-contact cases ever diverges in a non-source action. Six contacts increase this to four cases; their first divergences occur at steps 46, 79, 101, and 104, after the controlled prefix. The

post-hoc population-paired six-minus-one difference is $+0.00231$ $[+0.00010, +0.00498]$ for action-divergence rate and $+0.00398$ $[+0.00028, +0.00900]$ for late state Hamming. The probability of any diverging case increases by $+0.0625$ $[-0.0208, +0.1458]$, which remains imprecise. At 12 contacts, action divergence is concentrated in fewer populations and the late directional share is null; source persistence itself is lower than at one contact by -0.255 $[-0.376, -0.128]$. The evidence therefore supports high attenuation with some dose-sensitive leakage, not a monotone cascade law.

The formal one-, six-, and exploratory twelve-contact JSONL files contain 23,040 decisions each. Each released package includes its full SHA-256 manifest, protocol, cases, code, model/tokenizer provenance, and pre-output analysis specification.

M STAGE 10D MISTRAL SOURCE–TRANSMISSION REPLICATION

Before Stage 10B completed and before any Stage 10D prompt was evaluated, we froze a cross-family paired-branch assay using Mistral-7B-Instruct-v0.3 revision `c170c708c41dac9275d15a8fff4eca08d52bab71`. All 64 Stage 10B populations contribute two outcome-blind, hash-ranked source snapshots at exposure 32 (128 cases). Control and counter-history branches share the full 12-agent checkpoint and future schedule; only the source's first 27 visible events have both calibrated names swapped. The last three events, exposure, order, and event-wise agreement sequence are preserved. The controlled prefix contains six source–direct, six direct–two-hop, and six two-hop–probe contacts, followed by a frozen continuation to 120 steps.

Table 19: Prospective Mistral source–transmission profile. Intervals resample the 64 source populations after averaging their two cases.

endpoint	estimate [95% CI]
source counter-name action, first contact	$+0.094$ $[-0.008, +0.195]$
direct action after source block	$+0.039$ $[0.000, +0.086]$
cases with any non-source action divergence	0.844 $[0.781, 0.898]$
non-source action-divergence rate	0.01766 $[0.01532, 0.02012]$
final non-source memory reach	0.262 $[0.241, 0.284]$
late non-source state Hamming	0.0230 $[0.0186, 0.0276]$
group-share AUC	-0.00062 $[-0.00294, +0.00163]$
late group-share difference	-0.00018 $[-0.00449, +0.00425]$

The immediate source estimate is positive but its lower confidence bound crosses zero, so the frozen source-localization gate fails. Propagation nevertheless differs sharply from Qwen: most paired cases eventually produce a non-source action difference, although only 1.77% of non-source decisions differ and population-share contrasts remain near zero. The intervention therefore changes downstream trajectories without inducing a directional population takeover. Source response, action leakage, and macro displacement are distinct diagnostic coordinates, and their relation is model-dependent. The complete 61,440-decision JSONL has SHA-256 `6c237dee540d34431a328d97fe7c2d760471489ff6663bd70ab033dc42b71310`; the archive SHA-256 is `418a56f28ff69c6ec556eb81785390d57cfa696bc2898ebf03412ec16da8eecf`.

N STAGE 10E STOCHASTIC-DECODER SCOPE CHECK

Before any Stage 10E response, we fixed Mistral-7B-Instruct-v0.3, $T = 0.7$ categorical sampling over the two allowed tokens, decision-keyed random draws, the two Stage 10B population schedules, 64 populations, and the same estimands and gates. Keyed draws leave pair schedules, neighbourhood rotations, and option-order streams unchanged by decoding outcomes or micro-batching. Both schedule hashes match Stage 10B, all 51,200 decisions parse, and 5,000 bootstraps resample populations.

The long-horizon early–late contrast remains positive and the future placebo remains null, but the primary local slope interval includes zero; its strict gate therefore fails. Conditional switching falls from 0.617 at exposure 1–5 to 0.460 at exposure 26–30. This supports

Table 20: Frozen Stage 10E temperature-0.7 results.

endpoint	estimate [95% CI]
exact same-round conditional exposure slope	-0.00256 $[-0.00583, +0.00070]$
conditional switch, $e = 1$–5 minus $e = 26$–30	$+0.1509$ $[+0.1269, +0.1758]$
future-exposure same-round placebo	-0.00010 $[-0.00338, +0.00305]$

exposure-indexed decline under stochastic decoding without establishing decoder-invariant local gradients. The complete archive SHA-256 is `d7ec2c2d87736d95106ba000e7c1b7e3f0daa05cd8ab1acf86336c1bcdad7d11`.

O STAGE 10F COMPLEMENTARY-ROLE SCOPE CHECK

We prospectively changed the interaction objective rather than the population scaffold. Two agents independently choose one of two calibrated roles; different roles earn $+100$, while a role collision earns -50. Plasticity is switching after the agent's preceding collision. Before population output, we fixed the index-free $H = 30$ renderer, rotating neighbourhoods, two independent 32-population Mistral cohorts, three estimands, and both gates. Thus success requires differentiation rather than the shared-name convention studied in the main task.

All 51,200 Mistral decisions from 64 populations parse. The exact-same-round conditional exposure slope is -0.00537 $[-0.00801, -0.00277]$ across 5,849 unequal-exposure pairs, passing the frozen negative-CI gate. Conditional switching falls from 0.678 at exposure 1–5 to 0.235 at exposure 26–30; the cluster-paired early-minus-late contrast is $+0.4418$ $[+0.4174, +0.4660]$. The future-exposure placebo is -0.00196 $[-0.00461, +0.00076]$ and passes its contains-zero gate.

An independently frozen Qwen scope assay also contains 64 populations and 51,200 parsed decisions. Conditional switching falls from 0.731 to 0.230, and its early-minus-late contrast is $+0.5021$ $[+0.4792, +0.5245]$. The sharper same-round slope is -0.00309 $[-0.00674, +0.00053]$, so the negative-CI gate fails, while the future placebo $+0.00058$ $[-0.00353, +0.00479]$ passes. The cross-model transfer is therefore strongest for the long-horizon decline, not for invariant local gradients. These assays transfer the diagnostic signature across opposite social objectives, not to an unrestricted class of tasks. The Mistral and Qwen archive SHA-256 values are `3422875e3e0315dcfc274a4c2b2a6b280543dae87658e05b1be92a552959e82a` and `d28709c86c454e372c4981792a413631f87aa248137ebccbab93610000771b4b`.

P STAGES 10H–10K WINDOW-PHASE DIAGNOSTIC

Stage 10H prospectively varies the visible-history window in the Qwen complementary-role task. The $H \in \{10, 30, 60\}$ arms use 32 matched population schedules and 200/600/1,200 rounds, giving 128,000 parsed decisions. The frozen primary asks whether conditional switching continues to decline after the window fills: the pooled exact-same-round slope over $1 \le e/H \le 1.5$ is -0.0398 $[-0.1834, +0.0932]$, so its negative-CI gate fails; the future-exposure placebo $+0.0180$ $[-0.1260, +0.1647]$ passes.

The frozen curves reveal a different, sharply aligned structure. Conditional switching rises immediately after old events begin leaving the window: post-fill minus pre-fill contrasts are $+0.1465$ $[+0.1062, +0.1880]$, $+0.3251$ $[+0.2924, +0.3566]$, and $+0.3712$ $[+0.3568, +0.3851]$ for $H = 10, 30, 60$. Their matched-population mean is $+0.2809$ $[+0.2630, +0.3005]$. Because this breakpoint was discovered in the Qwen outputs, we label these intervals exploratory. A stricter post-hoc closed-loop difference-in-differences subtracts a longer-window arm's change over the same absolute-exposure bins: $H10 - H30$ at $e = 10$ is $+0.1657$ $[+0.1117, +0.2184]$, and $H30 - H60$ at $e = 30$ is $+0.3911$ $[+0.3561, +0.4270]$. Thus the shorter-window rebound exceeds a contemporaneous longer-window change; it is not merely an unadjusted bend in one curve.

The same analysis applies without new inference to the previously archived Stage 7B Qwen naming-game window arms (32 matched schedules, 46,080 decisions per arm). Despite a differ-

ent success objective and conditioning event, the $H10 - H30$ threshold difference at $e = 10$ is $+0.0761$ $[+0.0303, +0.1261]$ (22 supported populations), and $H30 - H60$ at $e = 30$ is $+0.0514$ $[+0.0265, +0.0803]$ (20 supported populations). This independent post-hoc task panel repeats the turnover-aligned direction; its sparse conditioned support and retrospective status are retained.

Before inspecting any Mistral outcome, we froze the same normalized bins and the $H30 - H60$ threshold contrast for cross-family confirmation. All 128,000 Mistral panel decisions parse, and the first 200 rounds of all three arms have identical schedule hashes. Table 21 reports the complete frozen gate set. The pooled post-fill slope and placebo pass, as do the pre-specified turnover rebound and $H30 - H60$ threshold contrast. Unlike Qwen, Mistral's $H = 10$ arm does not rebound (-0.0471 $[-0.0793, -0.0149]$); the cross-family confirmation is the longer-window breakpoint, not an invariant curve at every H.

Table 21: Matched-window role-task results. Qwen turnover was discovered post hoc; all Mistral rows were frozen before inspecting Mistral window outcomes.

estimand	Qwen	Mistral
pooled post-fill slope	-0.0398 $[-0.1834, +0.0932]$	-0.4431 $[-0.5564, -0.3142]$
future-exposure placebo	$+0.0180$ $[-0.1260, +0.1647]$	$+0.0559$ $[-0.1104, +0.2246]$
pooled turnover rebound	$+0.2809$ $[+0.2630, +0.3005]$	$+0.0618$ $[+0.0477, +0.0750]$
$H30 - H60$ threshold DiD	$+0.3911$ $[+0.3561, +0.4270]$	$+0.1352$ $[+0.1064, +0.1632]$

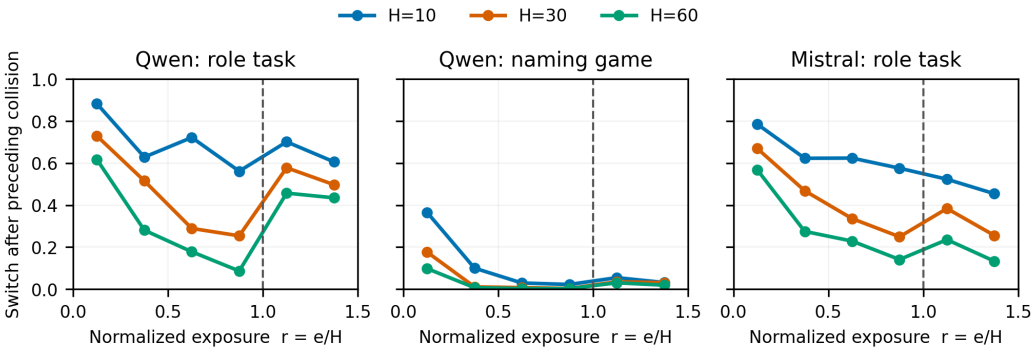

Figure 5: Conditional switching against normalized exposure $r = e/H$. The dashed line marks window fill. Qwen role-task curves motivated the turnover estimand; the archived naming panel repeats its direction post hoc; the Mistral role panel is the prospective cross-family test.

Stage 10K prospectively tests the discovered rebound in a 64-population, independent-seed Qwen $H = 30$ arm. All 76,800 decisions parse. The pre-specified turnover contrast is $+0.2863$ $[+0.2641, +0.3078]$, so its positive-interval gate passes. The long-horizon early-minus-late contrast is $+0.2242$ $[+0.2066, +0.2415]$. The post-fill slope -0.0022 $[-0.1919, +0.1497]$ is imprecise, while the future-exposure placebo -0.0020 $[-0.2230, +0.1773]$ contains zero. Thus the local threshold rebound—not a monotone post-fill slope—replicates under a new schedule seed. The complete archive SHA-256 is `9781a0bad2bfc78f2f6781e56cef888a3c7ea932ab9d4fe7ff892ea2f6e77953`.