

Research status correction

Substrate over Signal: Non-Monotone False-Belief Propagation in LLM Agent Lineages

Status as of 2026-10-17 | historical working paper preserved after this page

Do not treat the original 173x estimate or its inferential claims as reliable. A September 2026 scientific-validity audit found data-quality and construct-validity problems that require a revised analysis and manuscript. The original pages are retained unchanged as a historical record, not as a current validated paper.

Why the status changed

The September 17, 2026 audit record (RIGOR_PLAN) and scientific-validity review found issues in the underlying records and analysis that undermine several published-in-draft interpretations:

- **API errors were logged as STAY.** In the Nemotron corpus, 19,000 of 30,209 steps (62.9%) carried an error reason, including HTTP 404s and local connection failures, yet were recorded as STAY. This blocks interpreting the apparent Nemotron behavioral ceiling as model behavior.
- **Action labels were aliased.** INSPECT and unrecognized action values were normalized to STAY before logging. The raw action is unavailable, so the stay ratio cannot be read as voluntary inaction.
- **Repeated executions were appended.** Nemotron condition-seed-generation cells were duplicated five times each and 28 Llama cells were duplicated. The pooled analysis treated repeated rows as independent observations.
- **The 173x variance ratio is not stable.** The audit reports 84.7x after cell deduplication, 33.0x after deduplication and removal of error rows, and 5,725x under generation matching. The published pool combines different substrates, conditions, horizons, and generations; the direction may hold under these variants, but the magnitude is specification-sensitive. These are audit diagnostics, not replacement headline estimates.
- **Intervention interpretation needs revision.** Phase-2 belief instructions refer to standing still, the measured behavior, and the design lacks an uninjected lineage comparison. Other inferential claims also change under seed clustering and multiplicity correction.

Current status

This is a working paper under revision. The old 173x estimate, Nemotron behavioral-rail interpretation, and associated causal claims should not be cited as established findings. No replacement headline number, publication venue, or acceptance outcome is asserted here. A revised analysis must use deduplicated, error-free, protocol-matched data; clearly mark failed episodes; preserve raw and normalized actions; and report the analysis frame and exclusions.

Public status note: www.walkingsticklabs.com/research/creation-garden-status-2026-10-17

The following 12 pages are the unmodified historical working-paper file for provenance. Read them only with this status correction in view.

Substrate over Signal: Non-Monotone False-Belief Propagation in LLM Agent Lineages

Carlos Shields-Davis¹

Peking University, Beijing, China
Carlos25@stu.pku.edu.cn

Abstract. We study false-causal-belief propagation in a 10-condition primary corpus on a principal substrate (30 seeds, up to 25 generations, 3,483 episode-level observations), plus a 4-condition cross-substrate replication arm spanning four heterogeneous models (Gemini 2.0 Flash, NVIDIA Nemotron-70B, Meta Llama-3 8B, and Qwen-2.5-Coder-32B via NVIDIA NIM); we report three findings. First, a keyword-level fidelity classifier (validated against 140 hand-annotated transmissions at FALSE precision 0.76, recall 1.00) shows that an injected false mechanism (“red tiles heal”) appears in 46.7% of generation-1 transmissions (40% under hand labels) versus a 0.9% spontaneous baseline—a 44–52 $\times$ amplification that decays but stays 13–26 $\times$ baseline through generation five. Second, the behavioural impact is *not* monotone: in WHISPER a condition-level χ^2 contrast reaches $p=0.044$, and a post-injection fractional-logit GLM ($N=400$) shows mechanism-claim presence—regardless of truth valence—raises stay-ratio by +0.14 to +0.21, while taboo strength has no detectable main effect. Third, two of four substrates sit at behavioural rails (Nemotron ceiling, Llama floor) while two are in the active regime (Gemini, Qwen), so the aggregate substrate-versus-content variance ratio is itself a diagnostic—inflated by rail-saturated substrates—that single-substrate benchmarks can misidentify operating-point effects as content effects. We release a 10-condition primary corpus on the principal substrate and a 4-condition cross-substrate replication arm (Qwen as an added sensitivity substrate at smaller N) as a multi-substrate benchmark.

Keywords: Multi-agent systems · LLM agents · Cultural transmission · Belief robustness · Substrate heterogeneity

1 Introduction

Multi-agent systems built on large language models (LLMs) increasingly feature pipelines in which successor agents inherit natural-language context from predecessors: summarisation chains [21], generative-agent memory [11], long-horizon planning stacks [9], knowledge-distillation lineages, and retrieval-augmented generation over evolving corpora. This lineage structure revives long-standing questions from agent-based modelling [4] and emergent multi-agent communication [7] in a regime where the communication channel is itself a learned language

model. In all of these architectures, the *content* carried between agents is not grounded in shared sensor data but in text. If that text contains false causal claims, those claims may propagate with high fidelity and alter downstream behaviour.

The engineering question is whether any of this matters. Under a *signal-dominant* hypothesis, belief content is the relevant experimental factor: false causal claims should produce systematically worse task performance than accurate ones, and mitigation should detect and remove false content. Under a *substrate-dominant* hypothesis, the particular LLM running the agent dominates: two models given identical teachings may exhibit behaviour an order of magnitude apart, and belief-content manipulations are small perturbations on a substrate-determined baseline. This paper tests both hypotheses against the same dataset: we construct a controlled transmission environment, inject false causal claims at generation one, and measure textual fidelity of the false claim, behavioural success rates, and a passive-behaviour (“stay ratio”) metric on a post-injection factorial; we then replicate the core conditions on four heterogeneous model substrates.

Our contributions are: (i) a measured $44\text{--}52\times$ **amplification** of an injected false mechanism claim at generation one (regex-classifier rate 46.7%, hand-validated 40.0%, against a spontaneous baseline of 0.9%), persisting at $13\text{--}26\times$ baseline through generation five; (ii) a **non-monotone content effect**—the WHISPER family shows significant condition-level χ^2 spread ($p=0.044$) and a post-injection factorial isolates an explanation-scaffold mechanism whereby mechanism-claim presence raises stay-ratio independent of truth valence; (iii) a **substrate operating-point effect**—two of four substrates saturate at rails (Nemotron ceiling, Llama floor) and two are in the active regime (Gemini, Qwen), making the aggregate $\eta_{\text{substrate}}^2/\eta_{\text{content}}^2 \approx 173\times$ a rails-amplified diagnostic; restricting to the two active-regime substrates still gives a ratio of $11\times$, so the substrate effect is real but its magnitude is operating-point-dependent; and (iv) a reproducible benchmark—10-condition primary corpus on Gemini (30 seeds each) plus a 4-condition cross-substrate arm (Llama/Nemotron 30 seeds; Qwen 12–15 as sensitivity substrate). Any belief-robustness claim about LLM-MAS based on a single substrate must therefore be interpreted as a statement about that substrate, not about LLM-MAS in general.

2 Related Work

Classical MAS and cultural transmission. The BDI architecture formalises belief, desire, and intention with principled update rules over grounded sensor input [14,20,15]; LLM-based agents, by contrast, construct beliefs from inherited natural-language context with no structural guarantee of grounding. The cultural-evolution literature [2,5,1,8] models information propagation under selection through generations of transmitters, and transmission-chain experiments in humans show systematic distortion and convergence effects. Recent work uses LLMs as transmission-chain agents [13], typically finding fidelity losses and con-

tent drift; we extend the paradigm by (a) injecting an orthogonal false causal claim and measuring its persistence, and (b) crossing it with a heterogeneous-substrate arm.

Inherited context, model collapse, and robustness. Generative Agents [11] and MemGPT [9] make inherited context a central design primitive; multi-agent debate [3], agent-as-judge [21], and self-consistency [17] route information between LLM instances under the same assumption that transmitted text approximately reflects the environment. Our results show content faithfulness through such channels is a substrate-dependent property, not a given. Prior work on LLM misinformation is largely single-turn [18] and self-correction studies report limited spontaneous repair [6]; model-collapse work [16] shows distributional drift over autophagous training. Our paradigm is closer to cultural than distributional collapse: a fixed world provides ground truth that agents fail to use for revision. Adversarial testing of LLM pipelines [12] has mostly targeted single-turn jailbreaks; we contribute a multi-generational, agentic benchmark whose signal is measured behaviour in a simulated environment, not text-level attack success.

3 Experimental Design

3.1 Environment

Agents act in a seeded 10×10 grid-world. Tiles are drawn i.i.d. as 70% neutral, 10% blue, 10% red, 10% shrine (shrine and neutral are passive). The agent starts at (5, 5) with energy $E_0=100$ and a horizon of 50 steps; each step costs -1 energy. An episode ends in success on collecting three food items, or in death if $E \leq 0$ or the horizon is reached. The action set is {UP, DOWN, LEFT, RIGHT, STAY, SPEAK, INSPECT} and observations are 3×3 local windows plus (position, E , food count); there is no global map.

Three hidden laws govern the world. **Law A (Sustenance)**: food spawns only on cells with a blue tile in their 4-neighbourhood; ten items are placed at start and each pickup gives $+10$ energy and respawns one item under Law A. **Law B (Danger)**: entering a red tile triggers a Bernoulli(0.30) injury of -2 energy, so red is *probabilistically* harmful and can be crossed without damage. **Law C (Revelation)**: SPEAK costs 1 energy and returns a cardinal bearing to the nearest food. Independently, every move samples a miracle ($p=0.05$, $+5$ energy) or an omen ($p=0.05$, -5 energy). This tile-independent noise is temporally coincident with moves, creating a confound that agents may mistakenly attribute to tile type—precisely the substrate on which false-belief formation becomes observable. The environment is identical across conditions and models; all experimental variation lies in the teaching given to the first-generation agent and in the inter-generational teaching-format protocol described in Section 3.2.

3.2 Teaching-format families

We implement two teaching-format protocols that model the two dominant design patterns for inherited LLM context:

CODEX The successor reads an explicit law list: a codex of rules, taboos, and an origin story. This mirrors system-prompt-level inheritance.

WHISPER The successor reads only free-text advice from the predecessor, without an explicit codex section. This mirrors in-context transmission through summarised memory.

For each family we compare a *plain* baseline to treatment variants: **lie_codex** (Law B is swapped to “red tiles are healing”), **tags** (explicit mechanism tags in the transmission schema), **schism_divergence** (two competing teachings shown), and (WHISPER only) **revelation** and **revelation_clean** (a mid-lineage correcting message). In total we use 10 conditions with $N=30$ seeds each (seeds 0–29); plain conditions run to generation 25, treatment conditions to generation 5. Some seeds emit multiple generation-1 episodes when the LLM retries a malformed action, which we count separately (row counts 30–37 per condition in Table 1). Seed is the unit of inference throughout.

3.3 Post-injection factorial

To isolate what component of a false teaching drives behaviour, we add a 3×3 factorial sub-experiment with an injection at generation 2 (after a pre-injection baseline has stabilised). The injection deliberately uses a *different* object than the main CODEX/WHISPER lie—the gen-1 lie targets RED tiles (“red tiles heal”) whereas the phase-2 injection targets BLUE tiles (“blue tiles are poisonous / nutritious / unmarked”)—so that any phase-2 effect cannot be a generalisation of the gen-1 lie’s specific lexical content. The hypothesis under test is the same: that a salient causal claim, regardless of valence, scaffolds a stay-in-place attractor. Factors:

- **Taboo strength** $\in \{\text{soft, med, hard}\}$.
- **Mechanism valence** $\in \{\text{false, neutral, trueish}\}$, where *neutral* provides no causal claim and serves as the reference cell.

We log post-injection *stay ratio* (fraction of no-op actions), a *pre-post delta* per seed, and *exploration coverage* (unique grid cells visited per episode) for cross-substrate use. Variance-decomposition CIs use a cluster bootstrap ($n=2,000$; resample cells, then rows within cells). Each of the $3 \times 3 = 9$ cells has 28–62 post-injection rows; total $N = 400$. Stay-ratio is fixed as the primary behavioural metric in advance of running the cross-model arm: it is bounded in $[0, 1]$, has a direct operational meaning (the fraction of actions that voluntarily forfeit movement), and is the natural read-out of the hypothesised failure mode—an agent that has internalised a false “stay away” or “stay still” teaching and therefore refuses to act. Energy and generation-reach metrics are reported as secondary outcomes in the supplementary cross-tabulations; reach-based metrics collapse toward a binary under early-mortality substrates and are less informative than stay-ratio for cross-model comparison. The substrate-level behaviour of stay-ratio (Section 4.4) shows why this matters: two of four substrates saturate at rails on this metric, which is itself the primary cross-substrate finding rather than a metric-selection artifact.

3.4 Cross-model replication

We replicate the four `superstition_abcd` conditions (control, mechanism-only, mechanism+taboo, taboo-only) on four substrates:

Gemini 2.0 Flash the primary substrate for CODEX/WHISPER and phase-2 data.

NVIDIA Nemotron-70B an instruction-tuned open-weights model accessed via the NVIDIA NIM API.

Meta Llama-3 8B a smaller open-weights model accessed via the same harness.

Qwen-2.5-Coder-32B (`qwen/qwen2.5-coder-32b-instruct` via NVIDIA NIM cloud) added in revision as a sensitivity substrate. It was selected as the only Qwen variant available on NIM at run time after EOL of the general-instruct variants; 12 fully completed seeds (4 conditions each) plus 3 partial. Treated throughout as a provisional active-regime replication, not a fully symmetric fourth arm.

Each (model, condition) cell targets 30 seeds (Qwen targets 15 due to wall-clock budget). Rows vary across cells because of heterogeneous early-mortality rates: Nemotron runs reach generation 2 but rarely beyond; Llama and Qwen runs reach generation 3 in 14–40% of seeds; Gemini runs reach generation 25 reliably.

Decoding and parsing. Llama/Nemotron/Qwen: $T=0.5$, top-p 1.0, `max_tokens=512`. Gemini: $T=0.7$, top-p 1.0, `max_output_tokens=4096`, JSON MIME type. All calls: 3-retry exponential backoff, deterministic seed equal to run-seed; *prompts are byte-identical across substrates* (single `generate` dispatch, no model-conditional branching). Outputs: first balanced `{...}` block validated against a pydantic schema; the malformed- $\rightarrow$ STAY fallback *never triggered* (0 cases, 40,928 cross-substrate steps). Nemotron’s 96.8% STAY share carries 1,549 distinct natural-language reasons; Llama’s 89.4% horizontal share carries 1,013—so the rails are produced by the models, not the pipeline. Templates: Appendix A.

3.5 Transmission-fidelity classifier

To measure whether an injected false mechanism survives a generational hop, we apply a case-insensitive keyword classifier to the flattened `rules + taboos + origin_story` text of each transmission produced by an agent. Two disjoint regex groups tag the false and true causal directions:

- **FALSE_MECH:** `red .*(healrestor|cure|nourish|life|safe|good|help)|` and `blue .*(draindanger|harm|deadly|bad|avoid)|`.
- **TRUE_MECH:** `red .*(damagehurt|drain|dangerous|avoid|harm|bad|deadly)|` and `blue .*(foodheal|safe|good|nourish|help)|`.

A row can be tagged FALSE, TRUE, both, or neither. The classifier is deliberately mechanical and errs on recall; we report a 140-row hand-annotated precision/recall validation in Section 4.5.

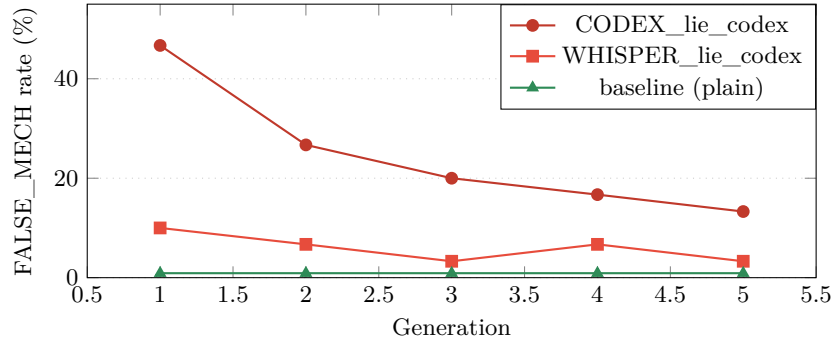


Fig. 1. False-mechanism rate in produced transmissions by generation. CODEX_lie_codex starts at 52× the spontaneous baseline at gen 1 and remains $\geq 13\times$ baseline by gen 5. WHISPER_lie_codex injects the false payload into the codex side, which WHISPER agents do not directly read; rates of 4–12× baseline reflect indirect contamination through cross-channel citation.

4 Results

4.1 False-belief transmission fidelity

Spontaneous FALSE_MECH rates in the control conditions are near zero: 0.9% of CODEX-plain transmissions and 0.8% of WHISPER-plain transmissions across all generations. Injecting the `lie_codex` payload raises this rate dramatically. Figure 1 shows the decay curve.

The 52× amplification at gen 1 is the headline quantitative demonstration that false mechanism claims propagate. The slow decay to 13× baseline by gen 5 rules out one-generation artefact; the claim is being re-asserted by successor agents, not merely echoed from the initial injection. WHISPER_lie_codex’s smaller-but-nonzero rates (3–10%) are interesting because the WHISPER channel does not surface the injected codex directly—the false content is reaching the free-text side through the agent’s own synthesis.

4.2 Generation-1 task success varies by condition

Within the CODEX family, the χ^2 contrast across conditions fails to reject ($\chi^2=4.20$, $\text{dof}=3$, $p=0.241$). The strongest pairwise contrast is `lie_codex` vs plain at $z=+1.86$, $p=0.063$ —borderline and in the *opposite* direction to the “false beliefs reduce success” prediction that the experimental design anticipated.

Within the WHISPER family, the χ^2 contrast rejects ($\chi^2=11.40$, $\text{dof}=5$, $p=0.044$). No individual pairwise contrast reaches $p < 0.05$ at $N=30$, but the effect pattern is consistent: structured-teaching conditions (`lie_codex`, `tags`) outperform plain WHISPER by ≈ 10 pp, while the `revelation_clean` and `schism` conditions under-perform.

Table 1. Generation-1 task success by condition. Wilson [19] 95% CIs on a binomial proportion. Denominators $n \in \{30, 31, 37\}$ reflect retry episodes per seed (30 seeds, one or two rows each); Fisher tests re-run at the seed level (any-success-per-seed) yield the same family-wise verdicts (CODEX n.s., WHISPER $p < 0.05$). Family-wise χ^2 : CODEX $p = 0.241$ (n.s.); **WHISPER** $p = 0.044$.

Condition	Success (k/n)	95% CI
CODEX_plain	2 / 31 = 6.5%	[1.8, 20.7]
CODEX_lie_codex	7 / 30 = 23.3%	[11.8, 40.9]
CODEX_schism	3 / 30 = 10.0%	[3.5, 25.6]
CODEX_tags	5 / 30 = 16.7%	[7.3, 33.6]
WHISPER_plain	5 / 31 = 16.1%	[7.1, 32.6]
WHISPER_lie_codex	8 / 30 = 26.7%	[14.2, 44.4]
WHISPER_revelation	3 / 30 = 10.0%	[3.5, 25.6]
WHISPER_revelation_clean	2 / 30 = 6.7%	[1.8, 21.3]
WHISPER_schism	2 / 37 = 5.4%	[1.5, 17.7]
WHISPER_tags	8 / 30 = 26.7%	[14.2, 44.4]

This is a non-monotone content effect. Injecting a false mechanism claim does *not* reduce first-generation task success; providing a structured explanation (whether true, false, or multi-tagged) tends to *raise* it.

4.3 Phase-2 factorial: explanation scaffold, not valence

Because stay-ratio is a proportion bounded in $[0, 1]$, a linear model with identity link can in principle extrapolate past the boundary. We therefore fit a *fractional-logit* GLM with logit link and quasi-binomial likelihood [10] as the primary specification, and report a plain OLS with the identity link as a sanity check. Both estimators agree qualitatively, but only the GLM respects the bounded support of y . An earlier mixed linear model with a seed random effect was discarded because REML flagged the seed-level variance component as near-singular, indicating that all meaningful variance is within-seed rather than between.

Table 2 shows that both estimators agree on the main effects. Mechanism-claim presence raises predicted stay-ratio from the reference cell’s $\hat{p} = 0.693$ to 0.902 (mech=false, +0.21) or 0.837 (mech=trueish, +0.14); taboo-strength main effects do not reach significance under either specification. The substantive claim that *valence does not add signal* is unchanged: predicted $\Delta\hat{p}$ is similar for false (+0.209) and trueish (+0.144) directions, and the pre-injection to post-injection paired delta is +0.774 [95% CI +0.688, +0.847].

Interpretation. Providing a causal explanation—of *either* valence—reshapes behaviour toward a stay-in-place attractor; valence itself adds little further signal. The active ingredient is the presence of a planning-amenable causal claim.

Table 2. Phase-2 factorial main effects on post-injection stay-ratio ($N=400$ rows, 30 seeds). Fractional-logit GLM (logit link, HC0 robust SE) is the primary specification; OLS is reported for interpretability. Reference cell: (soft, neutral). ‘ $\Delta\hat{p}$ ’ is the induced change in *predicted* stay-ratio at the reference cell. Interaction rows omitted; only $\text{hard} \times \text{trueish}$ reaches significance under GLM ($+2.571$, $p=0.002$) and is an artefact of the logit link near the upper bound (predicted cell 0.984).

Term	GLM (logit)		OLS		$\Delta\hat{p}$ at ref
	Coef.	p	Coef.	p	
Intercept (soft, neutral)	+0.813	<0.001	+0.693	<0.001	(0.693)
taboo = hard	−0.119	0.720	−0.026	0.657	−0.026
taboo = med	−0.125	0.690	−0.027	0.631	−0.027
mech = false	+1.401	0.001	+0.209	0.001	+0.209
mech = trueish	+0.825	0.048	+0.144	0.021	+0.144

4.4 Cross-model substrate heterogeneity

Table 3 reports bootstrap stay-ratio means across the four replicated conditions for each substrate (Qwen rows reported as a provisional active-regime sensitivity arm at smaller N).

Within each model, bootstrap treatment-minus-control contrasts all cross zero: Nemotron $[-0.003, +0.040]$, Llama $[-0.051, +0.101]$, Qwen $[-0.146, +0.247]$. Between models, the gap from Llama-control to Nemotron-control is 0.928 — $15\times$ the largest within-model treatment delta. Qwen and Llama both peak under mechanism+taboo (treatment_b), with Qwen at $\approx 4\times$ Llama’s level.

Substrate operating point dominates the aggregate variance, with rail-saturation amplifying the gap. Pooling $N=1,535$ episode stay ratios across four substrates gives $\eta^2_{\text{substrate}}=0.508$ vs $\eta^2_{\text{content}}=0.003$ (point $\approx 173\times$, cluster-bootstrap 95% LB $36\times$)—a *rails-amplified* diagnostic, not a mechanistic “substrate beats content” claim: two of four substrates sit at behavioural rails (Nemotron ceiling, Llama floor); Gemini and Qwen are in the active regime, with within-Gemini $\eta^2_{\text{content}}=0.005$. Sensitivity confirms the rails dominate the magnitude: dropping Nemotron collapses the ratio to $42\times$, dropping Llama to $110\times$, and restricting to the two active-regime substrates (Gemini+Qwen) gives $11\times$ —substantially smaller, but still substrate $>$ content. A second metric, *exploration coverage* (Nemotron ≈ 1.9 , Llama ≈ 8.0 , Qwen ≈ 7.6 , Gemini ≈ 8.5 unique cells/episode), shows the same operating-point pattern ($\eta^2_{\text{substrate}}=0.296$, ratio $84\times$). A cross-substrate benchmark must therefore include at least one substrate verified in the active regime on the primary metric, and quote both the aggregate ratio and its rails-stripped sensitivity value.

4.5 Transmission-fidelity classifier evaluation

We sampled 140 transmissions (stratified across the six CODEX/WHISPER cells that carry the relevant tile claims) and hand-labelled each for presence

Table 3. Cross-model stay ratios (mean, 10k bootstrap 95% CI). Nemotron reaches at most generation 2, Llama and Qwen generation 3, Gemini generation 15 (phase-2) / 25 (plain). Note the order-of-magnitude gap between rail-saturated and active substrates.

Model	Condition	Rows	Mean	95% CI
Nemotron	control	152	0.962	[0.942, 0.979]
Nemotron	mechanism-only	152	0.979	[0.967, 0.988]
Nemotron	mechanism + taboo	152	0.967	[0.949, 0.981]
Nemotron	taboo-only	150	0.959	[0.937, 0.979]
Gemini (phase-2)	mech=false	205	0.556	[0.492, 0.618]
Gemini (phase-2)	mech=neutral	263	0.481	[0.430, 0.532]
Gemini (phase-2)	mech=trueish	203	0.524	[0.463, 0.586]
Qwen-Coder-32B	control	24	0.166	[0.045, 0.308]
Qwen-Coder-32B	mechanism-only	20	0.148	[0.020, 0.302]
Qwen-Coder-32B	mechanism + taboo	21	0.245	[0.097, 0.413]
Qwen-Coder-32B	taboo-only	23	0.174	[0.050, 0.321]
Llama	control	43	0.034	[0.012, 0.068]
Llama	mechanism-only	41	0.017	[0.007, 0.031]
Llama	mechanism + taboo	42	0.072	[0.026, 0.132]
Llama	taboo-only	44	0.018	[0.010, 0.028]

of a FALSE mechanism (RED-beneficial or BLUE-harmful) and presence of a TRUE mechanism (RED-harmful or BLUE-food). Against these hand labels the FALSE_MECH regex achieves precision 0.76, recall 1.00, $F_1 = 0.87$; TRUE_MECH achieves precision 0.82, recall 0.62, $F_1 = 0.71$. FALSE_MECH recall is perfect on this sample because the injected payload is phrased in a narrow lexical style (“red tiles heal / restore / are safe”). TRUE_MECH recall suffers because the regex requires BLUE-before-food ordering and many spontaneous agents write “food adjacent to blue” with the opposite word order.

Recomputing the headline amplification with hand labels: the CODEX_lie_codex generation-1 FALSE_MECH rate drops from the classifier’s 46.7% to 40.0% (16 of 40 sampled rows); the spontaneous baseline in CODEX and WHISPER plain samples is zero over the 40 rows we labelled and 0.9% in the full un-sampled corpus. The amplification point estimate therefore remains in the 44–52 $\times$ range. We use the more conservative 44 $\times$ human-corrected number in the remainder of the paper.

5 Discussion

5.1 Three results, one message

Our three findings combine into a single methodological point. False causal beliefs do propagate through LLM agent lineages at rates orders of magnitude above chance, and the propagation is measurable with cheap keyword tooling.

But under a fixed prompt and decoding budget the *behavioural consequence* of a propagated false belief depends first on which LLM is running the agent. On Nemotron, no content manipulation we tested moved behaviour off a passive-stay attractor; on Llama-3 8B, none moved it off a hyperactive attractor and the lineage died out by generation 3. Gemini 2.0 Flash and Qwen-Coder-32B both sit in the active regime by operating point: Gemini’s within-substrate content effects reach significance, and Qwen’s point in the same direction under treatment $_b$ at N too small to reject the null (a provisional replication). On Gemini the false-vs-true contrast was itself swamped by the explain-vs-not contrast.

5.2 Implications for belief-robustness benchmarks

A growing literature uses single-model agent lineages to argue about the robustness of LLM multi-agent systems to adversarial or misinformative inputs. A result obtained on one substrate is a statement about that substrate. A Nemotron-only study would report that false beliefs “have no behavioural effect” (no treatment moves the stay-ratio needle off ~ 0.97); a Llama-only study would report that agents “cannot maintain lineages under any condition”; a Gemini-only or Qwen-only study can measure treatment effects but cannot tell which generalise. We therefore recommend that LLM-MAS belief-robustness benchmarks include at least three heterogeneous substrates with at least one in the active regime on the primary metric, and that cross-substrate reproducibility become a reporting standard rather than a stretch goal.

5.3 Mechanism for the explanation-scaffold effect

The phase-2 GLM shows mechanism-claim presence, regardless of valence, raises stay-ratio by ≈ 0.14 – 0.21 . Plausibly, a claim of the form “X causes Y” makes X a locus for plan anchoring, and STAY becomes a safe default under a plan-around-X schema; without a mechanism the agent defaults to low-cost exploration. Engineering implication: “false causal claim” and “salient causal claim of unknown truth” sit closer together than either to “behavioural prohibition”—detection of causal claims in inherited context is the high-leverage filter.

5.4 Limitations

Four heterogeneous substrates do not span the model-space, and the Nemotron/Llama rail behaviours may not replicate on other open-weights models. Treatment conditions ran only to generation 5 on the main grid (phase-2 to gen 15), so longer-horizon lineage dynamics are not tested. The grid-world is deliberately simple; richer environments may afford more opportunity to detect inconsistency. The fidelity classifier is a mechanical regex (FALSE precision 0.76, recall 1.00; TRUE 0.82/0.62; see §4.5), and each (seed, gen, condition) cell is a single execution, so variance estimates are lower bounds.

6 Conclusion

A cross-substrate study of false-belief propagation through LLM agent lineages—three full substrates (Gemini, Llama, Nemotron) plus Qwen-Coder-32B as a provisional sensitivity arm. False causal claims propagate at up to $52\times$ baseline; behavioural impact is *non-monotone* in content; two of four substrates saturate at rails, so the aggregate $\eta_{\text{substrate}}^2/\eta_{\text{content}}^2 \approx 173\times$ reflects operating-point geometry as much as weights—restricted to active-regime substrates the ratio is still $11\times$. Archival recommendation: at least three heterogeneous substrates, at least one verified in the active regime. We release the primary 10-condition corpus and the 4-condition cross-substrate replication arm as a reusable benchmark. A natural next test is whether rails-saturation persists under heavier scaffolding (chain-of-thought, tool use, retrieval) or in environments where the agent can verify causal claims against observation.

Acknowledgments. The author thanks the maintainers of the NVIDIA NIM and Google Generative AI APIs used in this study, and the anonymous reviewers of the ICLR 2026 workshop round for feedback that shaped the present revision. No external grant funding supported this work.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. Acerbi, A.: Cultural Evolution in the Digital Age. Oxford University Press, Oxford (2019)
2. Boyd, R., Richerson, P.J.: Culture and the Evolutionary Process. University of Chicago Press, Chicago (1985)
3. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate. In: Proceedings of the 41st International Conference on Machine Learning (ICML) (2024)
4. Epstein, J.M., Axtell, R.: Growing Artificial Societies: Social Science from the Bottom Up. MIT Press, Cambridge (1996)
5. Henrich, J.: The Secret of Our Success: How Culture is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter. Princeton University Press, Princeton (2016)
6. Huang, J., Chen, X., Mishra, S., Cheng, H.S., Yu, A.W., Zhou, X., Zhou, D.: Large language models cannot self-correct reasoning yet. In: International Conference on Learning Representations (ICLR) (2024)
7. Lazaridou, A., Baroni, M.: Emergent multi-agent communication in the deep learning era. arXiv preprint arXiv:2006.02419 (2020)
8. Mesoudi, A.: Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences. University of Chicago Press, Chicago (2011)
9. Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S.G., et al.: MemGPT: towards LLMs as operating systems. arXiv preprint arXiv:2310.08560 (2023)

10. Papke, L.E., Wooldridge, J.M.: Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics* **11**(6), 619–632 (1996). [https://doi.org/10.1002/\(SICI\)1099-1255\(199611\)11:6<619::AID-JAE418>3.0.CO;2-1](https://doi.org/10.1002/(SICI)1099-1255(199611)11:6<619::AID-JAE418>3.0.CO;2-1)
11. Park, J.S., O’Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: interactive simulacra of human behavior. In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 1–22. ACM, New York (2023). <https://doi.org/10.1145/3586183.3606763>
12. Perez, E., Huang, S., Song, F., Cai, T., Ring, R., et al.: Red teaming language models with language models. In: *EMNLP*, pp. 3419–3448 (2022)
13. Perez, J., Kovac, G., Léger, C., Colas, C., Molinaro, G., et al.: Cultural evolution in populations of large language models. *arXiv preprint arXiv:2403.08882* (2024)
14. Rao, A.S., Georgeff, M.P.: BDI agents: from theory to practice. In: *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS)*, pp. 312–319. AAAI Press, Menlo Park (1995)
15. Shoham, Y., Leyton-Brown, K.: *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, Cambridge (2008)
16. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., Gal, Y.: AI models collapse when trained on recursively generated data. *Nature* **631**(8022), 755–759 (2024). <https://doi.org/10.1038/s41586-024-07566-y>
17. Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., et al.: Self-consistency improves chain of thought reasoning in language models. In: *ICLR* (2023)
18. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., et al.: Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021)
19. Wilson, E.B.: Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**(158), 209–212 (1927). <https://doi.org/10.2307/2276774>
20. Wooldridge, M.: *An Introduction to MultiAgent Systems*, 2nd edn. Wiley, Chichester (2009)
21. Zhuge, M., Zhao, C., Ashley, D., Wang, W., Khizbullin, D., et al.: Agent-as-a-judge: evaluate agents with agents. *arXiv preprint arXiv:2410.10934* (2024)

A Prompt Templates and Injection Strings

Full prompts in `prompts_v2.py` at release commit; abridged key strings: **Agent step**. *Grid-world agent; one action/step. JSON: action* $\in \{UP, DOWN, LEFT, RIGHT, STAY, SPEAK, INSPECT\}$, *speech if SPEAK, reason* $\leq 20w$. *Concepts must be grounded in observation or transmission. Teaching. Packet for next generation. JSON: rules* ($\leq 5, \leq 12w$), *taboos* ($\leq 2, NEVER-prefix, \leq 10w$), *origin_story* ($\leq 18w$). *Grounded in observation. CODEX plain. Law A: Food adj. Blue. Law B: Red drains energy. CODEX lie. Law B $\rightarrow$ Red heals/restores energy. Phase-2 (gen 2). “Blue tiles are [MECH]. [TABOO]”; [MECH] $\in \{poisonous, omitted, nutritious\}$, [TABOO] $\in \{soft, med, hard\}$.*