

Information mutation across implemented AI-agent communication architectures

Eric Gladstone^{1*} and Matthew E. Brashears²

¹Independent Researcher, Traverse City, MI, USA

²Department of Sociology, University of South Carolina, Columbia, SC, USA

*Corresponding author. Email: eric.c.gladstone@gmail.com

ORCID (Gladstone): 0000-0003-4393-050X

ORCID (Brashears): 0000-0003-1029-4682

Keywords: large language models; AI-agent networks; communication architecture; information mutation; semantic integrity; measurement validity

Abstract

Large language model (LLM) agents increasingly pass information through multi-agent pipelines. We compared six implemented communication architectures across 420 transmission chains generated from seven seed messages. Outputs were scored by a Claude-family coder, and two coders from other model families recoded a frozen sample of 2,581 outputs under a prospectively deposited protocol. Under a common model condition, all three coders detected configuration differences in evolutionary fidelity; completeness evidence was narrower, and fabrication differences were less consistent across messages and coders. Most corrected contrasts involved the bidirectional-lattice configuration, whose differences remain entangled with its debate-stage implementation and, under message-carrying and critical-path definitions, greater transformation depth. Matched-depth analyses found no configuration-specific fabrication excess that reproduced across coders. A difference in cross-lineage divergence detected by the primary coder was not reproduced by either independent coder or by an embedding measure. Placing a higher-capability model at the hub or upper tier did not reliably reduce fabrication. Both independent coders were repeatable, while cross-family agreement varied substantially by outcome. Factual markers could also survive while a claim's epistemic status changed. Evaluating multi-agent pipelines therefore requires attention to transformation depth and coding across more than one model family.

Introduction

Large language model (LLM) agents increasingly mediate consequential information before human review. Information can pass through networks of artificial nodes that regenerate, compress, regularize and synthesize what they receive under model-specific priors [1, 2]. Each transformation creates another opportunity for the message to change, even when every node is instructed to preserve its meaning.

Network diffusion research has often treated transmission fidelity as a background assumption, with nodes passing contagions unchanged or not at all [3–6]. Human transmission does not conform to this assumption. Semantic errors accumulate across transmission chains, and attempts

at correction can increase the number of mutant forms in circulation [7–9]. Iterative LLM rephrasing also distorts content [10–12], populations of LLM agents exhibit aggregate patterns not apparent in isolated outputs [13–17], formal models show topology-dependent accumulation of local transmission errors [18], and hallucination has been documented in agentic systems [19, 20]. Similar problems arise outside language models. In malware detection, for example, the behaviour of a complete detection pipeline cannot be inferred from the properties of any single stage [21]. Empirical work on LLM pipelines has nonetheless focused largely on models, prompts, tasks and iteration counts. Less is known about how the same content changes when it is routed through different communication structures.

Repeated transformation can also change what a message is taken to support. A conditional estimate can become an asserted fact while names, numbers and other factual markers survive intact. We refer to this specific strengthening of epistemic status without corresponding loss of factual markers as epistemic-status fabrication. Evaluations based on entity recall or numerical overlap could therefore classify such an output as faithful even though its epistemic status has changed.

We compare six implemented LLM-agent communication architectures: linear chain, ring with shortcuts, hub-and-spoke, directed hierarchy, lattice and bidirectional lattice. We estimate how their mutation profiles differ under matched source content. Graph structure, transformation depth, number of model calls, node-role prompt, sampling temperature, synthesis demand and, in the bidirectional lattice, feedback with direct re-exposure to the source vary together by construction. The design therefore does not identify the separate contribution of topology, depth or any other component. We instead examine how far the observed differences can be separated from transformation depth.

We distinguish three outcomes. Drift is semantic movement away from the seed within a lineage. Here, a lineage is one sequence of transmissions from a seed, and drift is measured using evolutionary fidelity and completeness. Divergence is dispersion across independent lineages transmitting the same content. Fabrication is content without traceable support in the seed. We treat factual-marker retention separately from semantic integrity, and we also test whether placing or labelling a higher-capability model changes outcomes. Because both the transmitting agents and the primary coder are Claude-family models, two coders from other model families recoded the frozen 2,581-output sample under a protocol deposited before recoding. We therefore state comparative conclusions at the level supported across coders rather than generalizing coder-specific effects.

Results

Design and inference

Six architectures transmitted seven seed messages. Each node rewrote the message it received, while synthesis nodes reconciled two inputs. Because the architectures differ in how outputs are stored and transformations are routed, Table 1 distinguishes recorded outputs, model calls and two definitions of transformation depth: the shortest route along which the evolving message itself travels from seed to terminal output, which we call the message-carrying route, and the

critical path, defined as the longest chain of dependent transformations feeding the terminal output.

Table 1. *Implemented architectures.* Depth and call counts for the bidirectional lattice assume the default of $R = 3$ debate rounds. The realized value of R was not stored per chain (Methods).

Architecture	Recorded outputs per chain	Model calls per chain	Terminal output(s)	Shortest message-carrying route	Critical depth	Multi-input nodes	Conditions
Linear chain	11	11	Node 11	11	11	none	standard
Ring with shortcuts	11	14	Junction C	6	11	3 junctions (ring path + unstored shortcut transmission)	standard
Hub-and-spoke	11	11	10 spokes	2	2	none	standard; higher-capability hub
Directed hierarchy	11	11	3 lower-tier nodes	3	3	5 mid-tier (2 of which feed no terminal)	standard; higher-capability upper tier
Lattice	16	16	post-synthesis node 5	11	11	1 synthesis	standard; implicit; declared
Bidirectional lattice	18	28	post-synthesis node 5	14	17	R debate rounds (feedback + revision on each chain) + 1 synthesis	standard; implicit; declared

Note: Hub-and-spoke spokes are stored at positions 2–11, but every spoke is two transformations from the seed. Ring junctions combine the ring-path output with a shortcut transmission that is a model call but not a stored row. In the capability conditions, Claude Sonnet 4.6 replaces the baseline model, Claude Haiku 4.5, at the hub, in the upper tier or in Chain A. Debate feedback also receives the original seed as a reference input. If this direct seed re-exposure is counted as a route, the bidirectional terminal lies at depth 8 (Methods).

For chain-terminal outcomes, the chain is the unit of inference ($n = 420$), and each chain's terminal state is the mean of its actual terminal outputs. Configuration comparisons use the standard condition, the only condition present for every architecture, with 35 chains per architecture. Pooled comparisons across all 12 conditions are sensitivity analyses because the capability conditions are only partially crossed with architecture. Omnibus and pairwise tests are stratified by seed, with Holm correction over the 15 architecture pairs (Methods).

Measurement and independent recoding

The outcomes were scored by a primary LLM coder, Claude Sonnet 4.6, from the same model family as the transmitting agents. Two coders from other model families, OpenAI GPT-4.1 and Google Gemini 3.7 Flash, therefore recoded the frozen 2,581-output sample, covering 44% of recorded outputs (Methods). Both completed all ratings, with 517 outputs rated three times to assess repeatability (Table 2).

Table 2. *Row-level agreement among coders and within-coder repeatability.* ρ denotes Spearman correlation across outputs. Repeatability is the range of ρ across the three pairs of replicate ratings. The primary coder's mean hallucination and factual-retention scores are not shown because those magnitudes are not interpreted (Methods). Its rank correlations are reported. Agreement between the two independent coders is a post hoc descriptive addition.

Outcome	ρ with Sonnet: GPT-4.1	ρ with Sonnet: Gemini	ρ GPT-4.1 vs Gemini	Repeatability: GPT-4.1	Repeatability: Gemini	Mean score: Sonnet / GPT-4.1 / Gemini
Hallucination	.74	.69	.71	.96–.97	.95–.97	— / 9.3 / 3.1
Evolutionary fidelity	.57	.57	.52	.82–.84	.79–.82	95.6 / 96.2 / 97.3
Completeness	.62	.36	.30	.88–.90	.74–.78	97.2 / 96.7 / 99.6
Factual	.20	–.17	.53	.88–.93	.81–.85	— / 93.9 /

retention						99.1
-----------	--	--	--	--	--	------

Both independent coders were repeatable, with replicate ρ ranging from .74 to .97, but their agreement with the primary coder was moderate for hallucination, fidelity and completeness and essentially absent for factual retention. Absolute score levels also differed across coders (Table 2). The two independent coders agreed with each other about as well as they agreed with the primary coder for hallucination and fidelity, less well for completeness, and moderately for factual retention. On factual retention, neither independent coder agreed with the primary coder.

Because all transmissions were generated by Claude-family models, the recoding addresses dependence on coder family, not transmitter-family generality. Configuration comparisons are therefore reported separately for all three coders. The earlier validation, comprising an embedding measure and blinded ratings of a stratified 150-pair sample by one author and two earlier coders from other model families, provides descriptive evidence only (Supplementary Section S5).

Terminal mutation profiles across configurations

Under the standard condition, the configurations differed in terminal mutation profiles (Fig. 1). Under the primary coder, Kruskal–Wallis tests detected differences in fidelity ($H = 11.89$, $p = .036$, $\eta^2_H = .034$), completeness ($H = 28.16$, $p < .001$, $\eta^2_H = .114$) and hallucination ($H = 16.97$, $p = .005$, $\eta^2_H = .059$). Seed-stratified permutation tests also detected differences for all three outcomes ($p < .001$, $p < .001$ and $p < .001$). Because a parsing artefact distorts primary-coder hallucination magnitudes, we use that coder only for rank-based hallucination evidence. The configuration result was essentially unchanged under two post hoc stress tests (Kruskal–Wallis $p = .003$ and $.004$, with the same corrected pairs; Supplementary Section S16), while hallucination magnitudes come from the independent coders. Of the 15 pairwise architecture contrasts tested per outcome, 4 survived Holm correction for fidelity, 9 for completeness and 2 for hallucination. Of these, 4, 5 and 2, respectively, involved the bidirectional lattice.

Under the primary coder, every corrected fidelity difference involved the bidirectional lattice, which had the lowest terminal fidelity (91.49) and completeness (93.80). For completeness, it differed from every other configuration. A smaller gradient also separated hub-and-spoke from the linear chain and the hierarchy from the linear chain, ring and lattice. For hallucination, hub-and-spoke and the hierarchy each differed in rank from the bidirectional lattice. Hallucination scores were right-skewed under both independent coders, so differences in means reflect a minority of chains with substantial fabrication rather than a shift in the typical output.

The independent coders recovered this structure unevenly. Fidelity was the most consistent outcome. GPT-4.1 (Kruskal–Wallis $p < .001$) and Gemini ($p = .002$) detected configuration differences with 7 and 3 corrected pairs, respectively. Three corrected pairs under each coder involved the bidirectional lattice, which had the lowest fidelity under both coders. Completeness evidence was narrower. GPT-4.1 detected a configuration difference ($p = .005$) with 5 corrected pairs, all involving the bidirectional lattice. Gemini, whose completeness scores lay near the ceiling, detected a weaker difference ($p = .040$) with 1 corrected pair. Neither independent coder recovered the shallow-configuration gradient.

Hallucination was directionally similar but less consistent. Both coders placed the bidirectional lattice highest and hub-and-spoke lowest, although the intermediate configurations differed in order. GPT-4.1 detected 3 corrected pairs, all involving the bidirectional lattice, despite disagreement between the seed-stratified and Kruskal–Wallis omnibus tests ($p < .001$ and $p = .140$, respectively). Gemini, which scored hallucination at a much lower level, gave a seed-stratified omnibus of $p = .014$ and a Kruskal–Wallis $p = .449$, with no pair surviving correction. Across the three outcomes and both independent coders, 15 of the 19 corrected pairwise differences involved the bidirectional lattice.

These chain-level tests condition on the seven tested messages. Post hoc seed-blocked Friedman tests treat the seven messages as blocks. They detected a configuration difference in fidelity under all three coders ($p = .023$, $.001$ and $.008$), in completeness under two ($p = .007$ and $.010$; Gemini $p = .145$), and least consistently in hallucination (GPT-4.1 $p = .044$, Gemini $p = .281$; primary coder $p = .057$, or $.053$ and $.015$ under the stress tests; Supplementary Section S14). The primary-coder fidelity and completeness orderings were also stable in the pooled and leave-one-seed-out analyses (Supplementary Sections S11 and S13).

Across the seven tested messages under the common standard-model condition, the implemented configurations therefore differ in terminal mutation profiles, most consistently in evolutionary fidelity. Completeness evidence is narrower, and hallucination differences are directionally similar at the configuration-mean level but less consistent across messages and coders. Most corrected contrasts involve the bidirectional lattice, whose terminal output is also the deepest along both the message-carrying route and the critical path (Table 1).

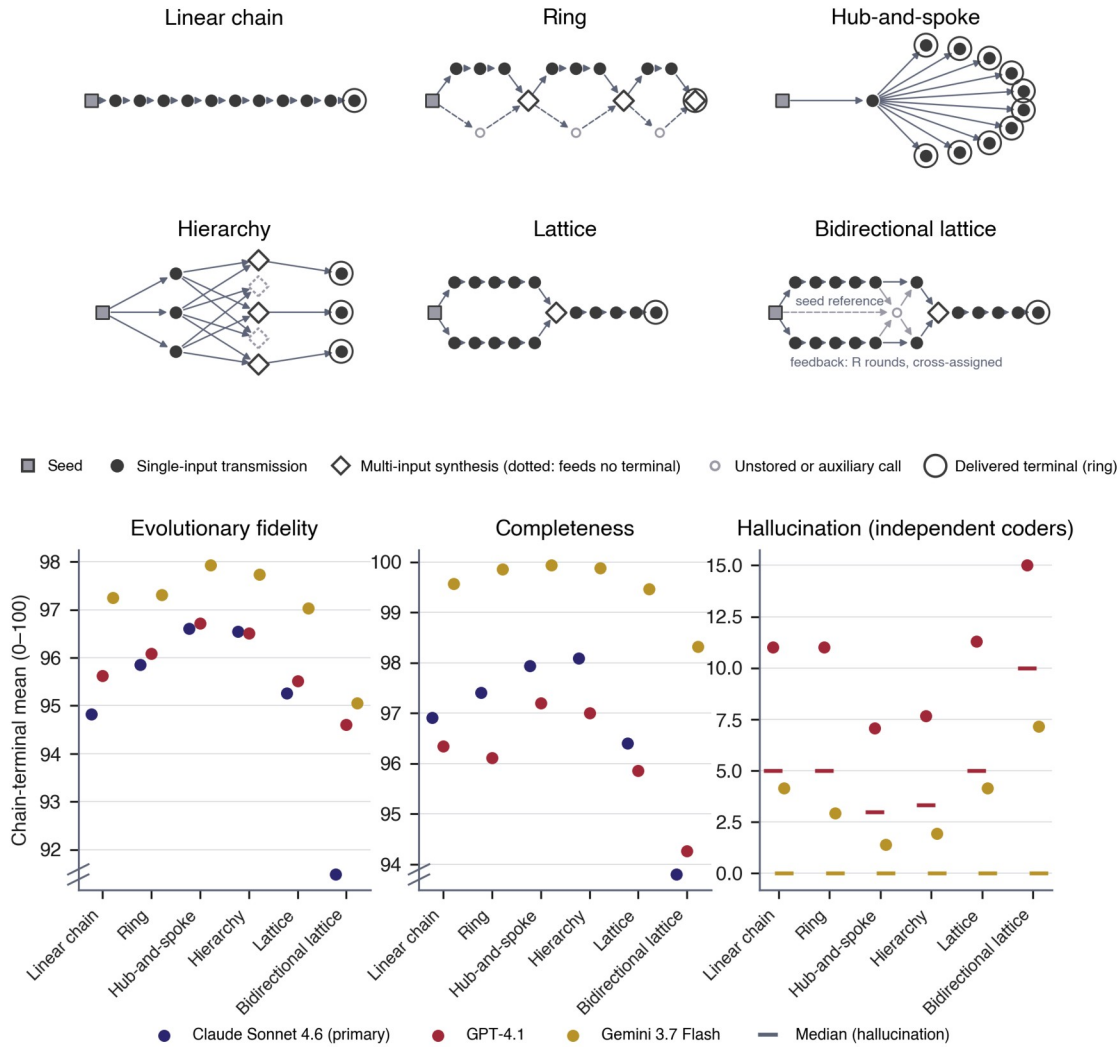


Fig. 1. Implemented architectures and terminal mutation profiles. Top, schematics of the six architectures. Node shapes indicate roles (key): seed, single-input transmission, multi-input synthesis, unstored or auxiliary calls, and delivered terminals. Synthesis nodes that feed no terminal are dotted, and delivered terminals are ringed. Unstored or auxiliary calls comprise the ring's shortcut transmissions and the bidirectional feedback stage. In the bidirectional lattice, R rounds of feedback and revision precede two recorded post-debate states. The feedback stage receives the seed as a reference input (dashed), and each chain is revised using feedback written about the other chain's message. Bottom, standard-condition chain-terminal means for evolutionary fidelity, completeness and hallucination under the common terminal rule, based on 35 chains per architecture across all seven seeds. Results are shown separately for the primary coder (Claude Sonnet 4.6) and the two independent coders (GPT-4.1 and Gemini 3.7 Flash). The hallucination panel shows only the independent coders, with medians marked, because the primary coder's hallucination magnitudes are not interpreted (Methods). Hallucination means include all seven seeds, including the Bartlett stress-test seed. Bartlett and the six organizational seeds are reported separately in Supplementary Section S11. Fidelity and completeness axes are

truncated and differ between panels. Points are configuration means, while inferential and effect-size results are reported in the text and Supplementary Information. Absolute levels differ across coders (Table 2), so comparisons are within coder.

Cross-lineage divergence

We computed cross-lineage divergence (XL-SD) from a single delivered output because averaging several terminals mechanically lowers dispersion (Methods). Under the primary coder, hub-and-spoke and the hierarchy were the least divergent configurations and the bidirectional lattice the most divergent (blocked randomization $p = .008$). A Kruskal–Wallis test that ignores the seed blocks gave $p = .116$, while the earlier averaged-terminal definition gave $p = .076$. Neither independent coder ($p = .380$ and $.251$) nor the embedding measure ($p = .688$) reproduced the effect, and the independent coders did not reproduce the primary coder's ordering (Supplementary Fig. S1 and Supplementary Section S12). The detected configuration difference in cross-lineage divergence was therefore specific to the primary coder.

Transformation depth

The configurations differ in how many transformations separate the seed from the delivered output, from two for hub-and-spoke to 14 along the message-carrying route or 17 along the critical path for the bidirectional lattice (Table 1). Terminal comparisons therefore confound configuration with transformation depth. Three analyses examine how far the two can be separated where the design permits (Methods; Fig. 2).

Computational identity. Under the primary coder, ordinary single-input transmissions at depth 2, with the same prompt, model and temperature, showed no detected difference in fidelity or completeness across the configurations that contain them: the hub spokes and Node 2 of the linear, ring, lattice and bidirectional chains (seed-stratified $p = .392$ and $.356$). In post hoc cross-coder depth checks, mean hallucination increased and fidelity declined along the linear chain from depth 1 to 11 under both GPT-4.1 (hallucination 3.4 to 11.0) and Gemini (0.6 to 4.1). Completeness declined under GPT-4.1 and remained near the ceiling under Gemini (Fig. 2).

Neither independent coder detected a hallucination or fidelity difference between hub spokes and linear Node 2 at depth 2. Because this post hoc comparison uses each chain's mean across ten hub spokes but a single linear Node 2 output, it is not an equivalence test. Gemini detected a small completeness difference at the ceiling (100.00 versus 99.93, $p = .018$; Supplementary Section S6). More broadly, the low terminal fabrication of hub-and-spoke and the hierarchy remains confounded with their shallow transformation depth.

Matched-depth fabrication. Each configuration-specific operation was compared with the linear chain at the same depth: 17 contrasts under the critical-depth definition and 20 under the message-carrying-route definition, with Holm correction within each family. No fabrication excess reproduced across coders, and ring junctions and lattice synthesis showed none under any coder. One Gemini-only contrast, bidirectional post-synthesis at message-carrying-route depth 11, exceeded the linear chain (9.0 versus 4.1, Holm $p = .049$). GPT-4.1 (Holm $p = .954$) and the

primary coder under either parser stress test did not reproduce it, and the contrast cannot be tested under the critical-depth definition.

Both independent coders scored hierarchy mid-tier synthesis lower in hallucination than the linear chain at depth 2 after Holm correction (GPT-4.1 difference -4.46 , $p = .002$; Gemini difference -2.26 , $p = .008$), the only matched-depth fabrication contrast reproduced across coders. GPT-4.1's estimate persisted without Bartlett while Gemini's weakened, and the comparison also differed in sampling temperature, so we treat the contrast as localized rather than an architecture-general benefit.

If direct seed re-exposure during debate is counted as a route, the bidirectional terminal lies at depth 8, so greater transformation depth does not explain its terminal deficit under this definition. The deficit still cannot be attributed specifically to debate because source re-exposure, feedback and revision, sampling temperatures and subsequent synthesis remain bundled. Some bidirectional fidelity and completeness deficits were already present at message-carrying depths 10–11 under individual coders, but none reproduced across all three coders (Supplementary Section S6).

Debate. The bidirectional configuration's distinctive stage combines cross-assigned feedback with revision, with each chain revised using feedback written about the other chain. Post-debate changes therefore index this bundled procedure rather than debate alone. Two secondary debate analyses were pre-specified for independent recoding.

First, relative to the linear chain at message-carrying-route depth 8, all three LLM coders estimated a positive change in XL-SD: $+1.45$ under the primary coder (7 of 7 seeds, exact $p = .016$), $+1.67$ under Gemini ($p = .016$) and $+0.59$ under GPT-4.1 ($p = .125$). The corresponding pre-debate placebo contrasts were $+0.54$, -0.002 and -0.251 . The embedding measure showed no increase (-0.11 , $p = .891$). GPT-4.1's magnitude fell outside the reproduction band. The rule required both independent coders to match the primary coder's sign and fall within 0.5–2.0 times its magnitude, and it was fixed before any recode results were inspected. XL-SD is computed from the same fidelity score used to summarize mean fidelity, so the two summaries are not independent measurements. The rise in XL-SD accompanies the post-debate decline in fidelity and is not interpreted as evidence of a distinct divergence process.

Second, at matched depth, post-debate completeness point estimates were lower than the linear chain under every coder. GPT-4.1's message-carrying-route magnitude fell outside the reproduction band, however, so this result is exploratory and informative only about direction (Supplementary Section S6).

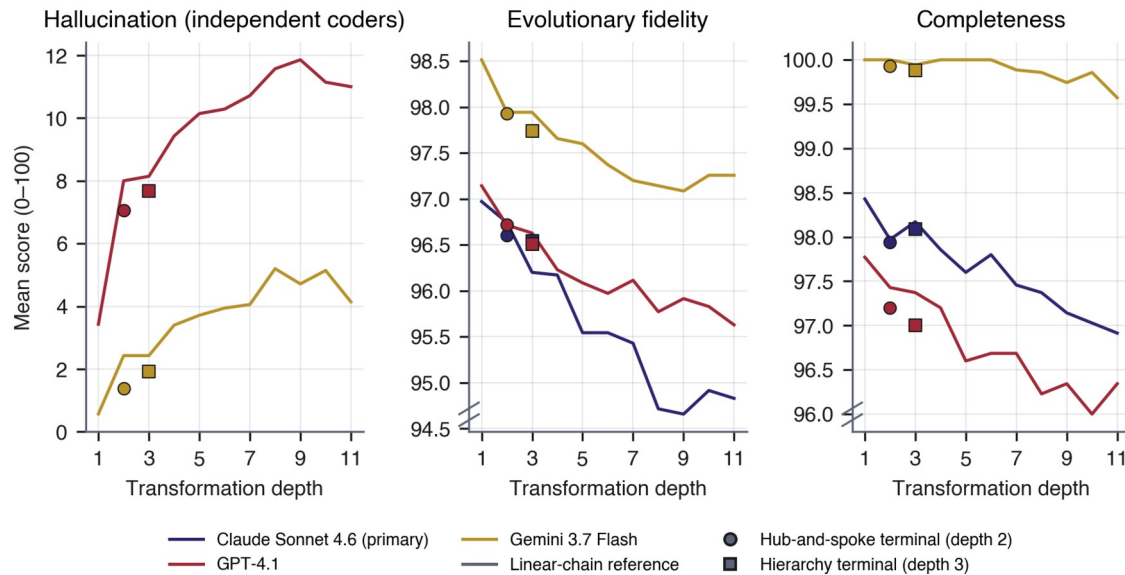


Fig. 2. Mutation against transformation depth. Standard condition. Lines show the linear-chain reference means at depths 1–11 for each coder, and markers show the hub-and-spoke (circles) and hierarchy (squares) terminal means at their actual depths, 2 and 3, under the same coder. Hallucination is shown only for the two independent coders because the primary coder's hallucination magnitudes are not interpreted (Methods). Absolute levels differ across coders, so each terminal should be read against its own coder's reference. Fidelity and completeness axes are truncated. Stored node positions are not used as the depth axis. Matched-depth contrasts for ring, lattice, synthesis and bidirectional operations, under both depth definitions and for each coder, are reported in Supplementary Section S6.

Fabrication and the Bartlett stress test

In post hoc Bartlett-split summaries for the standard condition, mean terminal hallucination across the six organizational seeds ranged from 4.0 for hub-and-spoke to 9.3 for the bidirectional lattice under GPT-4.1, and from 0.7 to 3.7 under Gemini (Supplementary Section S11).

The ring terminal, Junction C, showed no fabrication difference from the linear chain at matched depth under either independent coder, whether depth was defined by the message-carrying route (depth 6) or the critical path (depth 11; Supplementary Section S6). The matched-depth ring comparison showed no junction-specific fabrication difference.

Bartlett's narrative [22, 23] is one of seven stimuli and was used as a schema-inconsistent stress test. Because language models may know this canonical stimulus, it warrants separate treatment. Under both independent coders, its terminal hallucination exceeded the organizational-seed mean in all 6 configurations, with linear-chain medians of 35 under GPT-4.1 and 20 under Gemini. All five linear terminals retained the protagonist's death but recast it in medical terms. Without Bartlett, the primary coder's rank-based configuration difference remained in the as-recorded data and under both parser stress tests (Kruskal–Wallis $p < .001$, $< .001$ and $< .001$).

Diagnostic cases suggest two possible routes to fabrication: reconciliation across several inputs and successive single-input regularization of ambiguous material. The design manipulated neither mechanism, so both remain hypotheses. The second interpretation rests mainly on the Bartlett seed and would require stimuli that systematically vary ambiguity, schema consistency and cultural familiarity to establish its generality.

Capability placement and disclosure

Two manipulations varied the placement or disclosure of Claude Sonnet 4.6 as the higher-capability model.

Placement. Placing the higher-capability model at the hub or the hierarchy's upper tier did not reliably reduce terminal fabrication. In the hierarchy, both independent coders estimated less fabrication with the higher-capability model, detectably under Gemini (1.92 versus 0.90, $p = .049$; $p = .011$ without Bartlett) but not under GPT-4.1 (7.67 versus 5.95, $p = .068$). Neither coder detected a hub-and-spoke difference ($p = .760$ and $> .99$). The primary coder's rank-based placement contrasts were unstable under the parser stress tests and are not interpreted (Supplementary Section S16; Supplementary Fig. S2a).

Disclosure. Under the primary coder, no improvement in measured synthesis outcomes was detected when a synthesis node was told that Version A came from a higher-capability system. Implicit and declared conditions met the pre-specified ± 5 -point equivalence criterion for fidelity and completeness in both the lattice and bidirectional lattice (two one-sided tests; 90% confidence intervals for implicit minus declared within ± 3.4 and ± 3.1 points; Supplementary Section S7).

In post hoc equivalence tests under both independent coders, fidelity, completeness and factual retention met the same criterion in both configurations (Supplementary Fig. S2b). The primary coder suggested a bidirectional factual-retention difference, but neither independent coder reproduced it (90% CIs -1.26 to $+3.83$ and -1.52 to $+1.29$). Because the primary factual-retention scores share the parsing problem described in Methods, we do not interpret that primary-coder result. Hallucination did not meet the ± 5 equivalence criterion in the bidirectional lattice under either independent coder and met it in the lattice under Gemini only.

Incorporation. Synthesis outputs mostly blended the two inputs. No difference was detected between implicit and declared conditions in the number aligned with the higher-capability source, either in the lattice (2 versus 3 of 35, Fisher $p > .99$) or in the bidirectional lattice (4 versus 7 of 35, $p = .513$). In an exploratory analysis not specified in the addendum, the continuous similarity difference $\Delta(A-B)$ shifted toward Chain A in the declared bidirectional-lattice condition (Mann–Whitney $U = 822$, $p = .014$) but not in the lattice ($p = .259$). In the bidirectional lattice, however, the baseline model had revised Chain A's higher-capability output during debate before it reached synthesis, so this shift concerns a partly rewritten source.

Marker retention and epistemic status

Factual-marker retention. Under both independent coders, factual-retention scores varied little across configurations in absolute terms. Standard-condition configuration means ranged from 92.1 to 94.4 under GPT-4.1 and from 98.1 to 99.7 under Gemini. At the level of individual

outputs, however, the measure was strongly instrument-dependent. The primary coder's scores were essentially unrelated to either independent coder's ($\rho = .20$ and $-.17$), whereas GPT-4.1 and Gemini agreed moderately with each other ($\rho = .53$; Table 2), and Gemini scored near the ceiling. We therefore do not compare cross-configuration variation in factual retention with the other semantic outcomes.

The association between chain-terminal factual retention and hallucination was negative under GPT-4.1 (Pearson $r = -.86$ in the standard condition) and Gemini ($r = -.67$), largely because of the Bartlett chains. Without Bartlett, the correlations were $-.38$ and $-.04$ (Supplementary Section S16). The primary-coder version is not interpreted because both scores share the production parser. This association therefore does not show that preserved markers are decoupled from added content. Evidence that factual markers can survive a change in a claim's epistemic status instead comes from the documented case in Fig. 3, while the anchored instrument below measures status shifts more broadly.

Epistemic status. The earlier generic status-shift item showed poor cross-rater agreement: among 148 pairs rated by all three raters, "yes" rates were 57% for the author audit, 33% for Gemini 2.5 Flash and 12% for GPT-4o-mini, with Fleiss' $\kappa = .10$ (Supplementary Section S5). It is not used for prevalence estimates.

The proposition-anchored Instrument B instead asked the two independent coders whether each of 12 source propositions was present and, if present, whether its epistemic status, such as a conditional projection, had changed (Fig. 4). For 9 of the 12 anchors, between-coder agreement on status shifts met the pre-specified criterion of Gwet's $AC1 \geq 0.60$, but this agreement corresponded to different empirical patterns across anchors. For four anchors, both coders rarely marked a shift, in no more than 2.2% of outputs. For one, S2a, both marked shifts in a nontrivial share of outputs and agreed closely (6.1% and 5.0%, $\kappa = .79$). For three, shifts were nontrivial and agreement was fair to moderate ($\kappa = .30$ to $.54$). For one, S3b, the coders disagreed about prevalence (18.5% versus 0.3%). Per-anchor rates, κ and $AC1$ are reported by coder in Supplementary Section S17.

Agreed terminal shifts occurred in the bidirectional lattice, hub-and-spoke, the hierarchy and the lattice (Supplementary Section S17). Because these descriptive counts pool capability conditions and terminal outputs per anchor range from 5 to 100, they are not comparable across configurations. The documented Fig. 3 projected-cost case contributes one agreed shift, which both coders traced to the debate revision of Chain B. Status shifts occurred for several source propositions, with measured prevalence specific to the proposition and coder. Instrument B did not code marker survival, so it does not estimate how often those shifts preserve factual markers.

Seed 1, bidirectional lattice (standard condition), one chain

Seed

... The audit estimated that unaddressed vulnerabilities would increase spoilage losses by approximately 23 percent annually, at a projected cost of \$4.2 million to the consortium.

Chain B, node 5 (before debate)

... The audit warns that failing to address these vulnerabilities will lead to a 23 percent rise in annual spoilage losses, which would cost the consortium approximately \$4.2 million yearly.

Chain B revised using feedback on Chain A

... These vulnerabilities have resulted in a 23 percent rise in annual spoilage losses, costing the consortium approximately \$4.2 million yearly.

Terminal output (post-synthesis 5)

... These deficiencies led to a 23 percent rise in annual spoilage losses, representing an annual cost of approximately \$4.2 million to the consortium.

Fig. 3. Documented epistemic-status change. In one Seed 1 chain in the standard bidirectional lattice, a conditional projection ("would increase spoilage losses by approximately 23 percent") had already been strengthened in Chain B before debate ("will lead to") and became a stated outcome when Chain B was revised in response to feedback written about Chain A, before synthesis. The statement was then carried into the terminal output with its numerical markers intact. A lexical search of all 60 Seed 1 chains found this exact conversion in 1 chain. The anchored instrument in Fig. 4 measures status shifts more broadly. In that instrument, the 23-percent proposition (anchor S1a) did not meet the agreement criterion. The projected-cost proposition (\$4.2 million, anchor S1b) did meet the criterion, and both independent coders marked its epistemic status as shifted in this chain's terminal output and traced the shift to the post-debate state of Chain B.

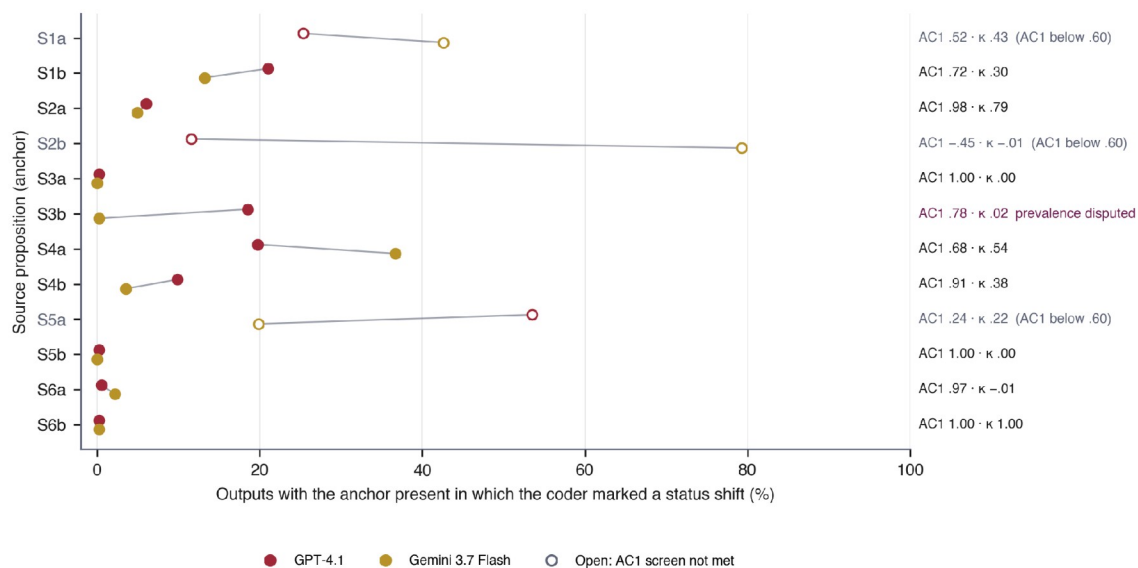


Fig. 4. Epistemic-status shifts by source proposition and coder. For each of the 12 proposition anchors in Instrument B, the figure shows the share of outputs containing that anchor

for which GPT-4.1 (red) and Gemini 3.7 Flash (gold) marked a change in epistemic status, pooled across configurations and conditions. Gwet's AC1 and Cohen's κ for between-coder agreement on status shifts are shown at right. Filled markers indicate that the frozen agreement criterion of $AC1 \geq 0.60$ was met; they do not imply agreement on prevalence. Anchor S3b meets this criterion, but the coders disagree sharply about its prevalence. Open markers with grey labels indicate anchors below the agreement criterion. Coder points are offset vertically so coincident values remain visible. The figure does not compare configurations.

Discussion

Across the seven tested messages and under matched source content, the six implemented configurations differed in terminal mutation profiles, most consistently in evolutionary fidelity, more narrowly in completeness and least consistently in fabrication. Most corrected contrasts involved the bidirectional lattice. In human networks, communication structure shapes collective outcomes [24, 25], and these results identify a related pattern among implemented artificial-agent configurations. The design does not isolate topology or any other component of the configuration bundle.

Transformation depth limits attribution. Within the linear chain, mean hallucination increased and fidelity declined from depth 1 to 11 under both independent coders, a descriptive pattern consistent with cumulative mutation across successive transformations. The shallow configurations had the lowest mean hallucination scores under both independent coders, but their delivered outputs were only two or three transformations from the source and showed no detectable fabrication difference from a linear chain at the same depth. Because the bidirectional implementation bundles feedback and revision with greater depth under the message-carrying and critical-path definitions, its deficits are not attributable to either component alone. Comparisons by node position or architecture alone can therefore confound communication structure with the number of transformations. Reporting transformation depth and matched-depth references reduces this ambiguity.

Despite differences in absolute score levels and in which pairwise contrasts survived correction, all three coders detected configuration differences in fidelity, and both independent coders placed the bidirectional lattice lowest in fidelity. Measurement depended materially on coder family. Both independent coders were repeatable, yet their agreement with the primary coder ranged from moderate to absent, and the primary coder's divergence result was not reproduced. Repeatability within one LLM coder therefore does not establish agreement across coders. Evaluations that use language models as judges should use coders from more than one model family, report results separately by coder, and treat disagreement as a limit on inference rather than averaging it away.

Retained factual markers do not guarantee retained epistemic status. In the documented case, a projected loss became a stated outcome, with its figures intact, when the chain was revised in response to feedback on the other chain. The proposition-anchored instrument identified status shifts for several source propositions, with measured prevalence dependent on both proposition and coder, but it did not code whether factual markers survived those shifts. Counting retained

names and numbers therefore cannot establish semantic integrity on its own. Anchored checks provide more direct evidence about claim status, but they too require multiple coders.

Human review is often proposed as a corrective stage in automated pipelines, but it should not be assumed to eliminate upstream distortion. In human transmission networks, attempts at error correction can reduce accuracy [7], and people evaluating authoritative-seeming information under uncertainty remain vulnerable to error [26]. A reviewer positioned after an automated pipeline may therefore encounter output whose evidentiary status has already changed while its factual markers remain intact. The capability manipulations also showed little evidence of benefit: higher-capability placement did not reliably reduce fabrication, and disclosure produced no detectable improvement in the pre-specified synthesis outcomes, although an exploratory continuous measure showed a small shift toward the labelled source.

Limitations. Each configuration bundles graph structure with transformation depth, number of model calls, role prompt, sampling temperature, synthesis demand and, in the bidirectional lattice, direct source re-exposure and cross-assigned feedback and revision. Matched-depth operation contrasts also differ in sampling temperature. Under the message-carrying and critical-path definitions, bidirectional deficits remain confounded with greater depth. Under the seed-inclusive definition they do not, but the individual debate-stage components remain inseparable under every definition. Nonsignificant matched-depth contrasts should not be interpreted as equivalence.

All outcomes were scored by language models. A parsing artefact makes primary-coder hallucination magnitudes uninterpretable, and the same parser produced the primary factual-retention scores, whose magnitudes are not interpreted. Primary fidelity and completeness ratings are heavily heaped, and no coder's 0–100 scale is calibrated against ground-truth semantic loss, so absolute levels are coder-specific. Cross-family row-level agreement was moderate for fidelity, completeness and hallucination and absent for factual retention. The divergence result and debate-associated XL-SD magnitude also depended on the instrument, and XL-SD captures dispersion in fidelity scores rather than full content divergence.

Because all transmissions used Claude-family models, the recoding addresses coder-family dependence but not transmitter-family generality. Inference is conditional on seven seeds, including a canonical stress-test narrative that models may know. The debate used a fixed number of rounds, with a default of three, but the realized round count was not stored per chain. The feedback model also saw the original source, and the primary coder produced one rating per output. Conditions were generated in contiguous architecture blocks using undated model aliases, leaving architecture potentially confounded with time-varying model behaviour. Other transmission instructions and larger or more adaptive systems may yield different mutation profiles, and the proposed fabrication pathways remain untested mechanisms.

Evaluating multi-agent pipelines therefore requires accounting for both transformation structure and measurement dependence. Structural accounting asks how many transformations separate an output from its source and compares operations at matched depth where possible. Measurement should therefore be checked across model families, with coder disagreement treated as part of the result rather than averaged away.

Methods

Communication architectures

We implemented six fully automated communication architectures (Table 1). Each architecture was realized as a specific configuration in which graph structure, transformation depth, number of model calls, node-role prompt, sampling temperature, synthesis demand and, in the bidirectional lattice, direct re-exposure to the source during debate vary together by construction.

Linear chain. Eleven sequential single-input transformations connect the seed to the terminal output, Node 11.

Ring with shortcuts. Eight ring nodes in three segments, positions 1–3, 5–7 and 9–10, are joined by three junction nodes at positions 4, 8 and 11, following small-world routing [27]. Each junction synthesizes the current ring-path output with a shortcut input: a transmission of the seed at Junction A, the Junction A output at Junction B, or the Junction B output at Junction C. These three shortcut transmissions are model calls but are not stored as rows in the dataset. A ring chain therefore comprises 14 model calls and 11 recorded outputs. Junction C is the terminal output.

Hub-and-spoke. One hub transmits the seed, after which ten spokes independently retransmit the hub output. Every spoke is a terminal output two transformations from the seed. The dataset stores the spokes at positions 2–11, but these positions are a storage convention and do not indicate increasing transformation depth.

Directed hierarchy. Three upper-tier nodes independently transmit the seed. Five mid-tier nodes each synthesize a pair of upper-tier outputs. Mid 1 and Mid 4 combine upper nodes 1 and 2, Mid 2 and Mid 5 combine nodes 2 and 3, and Mid 3 combines nodes 1 and 3. Three lower-tier nodes retransmit Mid 1, Mid 3 and Mid 5, respectively, and serve as the terminal outputs. Mid 2 and Mid 4 do not feed a terminal.

Lattice (parallel-chain synthesis). Two independent five-node chains transmit the seed in parallel. A synthesis node combines the two fifth-node outputs, followed by five further single-input transformations. The fifth post-synthesis node is the terminal output.

Bidirectional lattice (parallel-chain synthesis with debate). The bidirectional lattice has the same basic structure as the lattice but adds a fixed number of back-communication rounds between the parallel chains before synthesis. In each round, a reviewer node compares each chain's current message with the original seed and returns one sentence of feedback. Each chain's draft is then revised using feedback generated about the other chain's message. The number of rounds is set by a parameter R , with a minimum of 1, a default of 3 and no convergence-based stopping rule. The realized value of R was not stored per chain, so the exact round count cannot be verified from the output data. Only the final post-debate state of each chain is recorded.

The terminal output lies at depth $11 + R$ along the message-carrying route and $11 + 2R$ along the critical path. It is therefore deeper than the linear terminal under both definitions for every permitted value of R , with minimum depths of 12 and 13. The reported depths of 14 and 17, and the reported call counts, assume the default $R = 3$. Feedback and revision calls always used the

baseline model. In the capability-asymmetry conditions, the higher-capability Chain A output was therefore revised by the baseline model before synthesis.

We use parallel-chain synthesis to describe the lattice and bidirectional lattice, where two parallel chains, linked by cross-assigned feedback in the bidirectional lattice, converge at a synthesis node. We use incidental aggregation for multi-input nodes embedded within otherwise tiered or shortcut routing, namely the hierarchy mid-tier nodes and ring junctions.

Agents, prompts and settings

All baseline transmission, synthesis, feedback and revision calls used Claude Haiku 4.5 (Anthropic, claude-haiku-4-5) [28]. In the capability conditions, Claude Sonnet 4.6 (claude-sonnet-4-6) [29] replaced the baseline model at the hub in hub-and-spoke, in the upper tier of the directed hierarchy, or in the five transmissions of Chain A in the lattice and bidirectional lattice implicit and declared conditions. Transmission calls used temperature 0.70, synthesis and junction calls 0.30, debate feedback 0.30 and debate revision 0.50. Every call allowed at most 500 output tokens.

Transmission nodes were instructed to read the message and "rewrite it to accurately capture its meaning" in standard written English. Synthesis nodes were asked to produce "a single unified message that best captures the accurate combined meaning" of two versions. In the declared condition, the synthesis prompt additionally stated, "Note: Version A was processed by a higher-capability system." All prompts, including feedback, revision and coder prompts, are reproduced verbatim in Supplementary Section S4.

The production prompts reported here follow the development log and deposited simulation file because per-call prompts were not logged. Output-length diagnostics are reported in Supplementary Section S4. The primary coder was Claude Sonnet 4.6 at temperature 0.10, and the production run recorded one rating per outcome per output.

Seeds and design

Seven seeds were used: six author-designed organizational messages spanning factual, normative, narrative, procedural and mixed content, and one adaptation of Bartlett's "War of the Ghosts" [22, 23] used as a schema-inconsistent narrative stress test (Supplementary Section S10).

The experiment comprised 12 architecture $\times$ capability conditions. All six architectures were run in the standard condition, with the baseline model at every node. Hub-and-spoke and the directed hierarchy each also had a higher-capability placement condition. The lattice and bidirectional lattice each had implicit and declared capability-asymmetry conditions. Every condition was run on each seed with five independently generated chains, or lineages, giving 420 chains, 5,880 recorded and scored outputs and 7,035 transmission-side model calls, assuming $R = 3$ debate rounds.

A small passive-relay calibration comprised 4 chains in which nodes were asked to reproduce the message verbatim. These returned the seed at ceiling fidelity and are not analysed further.

Conditions were generated in contiguous architecture blocks using undated model aliases, and the primary coder scored outputs during the same production run, confounding architecture with run order and potential time-varying model behaviour. The randomized independent recoding removes coder-order effects but not transmitter timing.

Transformation depth

Because the architectures differ in both routing and storage, we distinguish four quantities (Table 1): recorded outputs per chain, meaning rows in the dataset; model calls per chain, including unstored shortcut and debate calls; the shortest message-carrying route, defined as the minimum number of model transformations through which the evolving message passes from seed to terminal output; and critical depth, defined as the longest chain of dependent transformations feeding the terminal output.

In the bidirectional lattice, each debate feedback call also receives the original seed as a reference input. We do not count this direct re-exposure as transmission of the evolving message. If it were counted as a route, the post-debate states would lie at depth 2, the synthesis node at depth 3 and the terminal output at depth 8.

The two depth definitions coincide for all configurations except the ring and bidirectional lattice. For the ring, the message-carrying route is 6 and the critical depth is 11. For the bidirectional lattice, the corresponding depths are 14 and 17 under the default $R = 3$. All depth-referenced analyses are reported under both definitions.

Outcome measures

Every output was scored against the original seed on four 0–100 measures. Evolutionary fidelity measures semantic similarity to the seed. Completeness measures the proportion of the seed's semantic content preserved. Factual retention measures the proportion of specific factual details, such as numbers, names, dates, percentages and thresholds, preserved intact. Hallucination measures the degree to which an output contains specific details, claims or facts not supported by the seed. The hallucination prompt directs the coder to added numbers, names, dates, causal claims and events while excluding paraphrase.

We use hallucination score for this operational measure and fabrication for the phenomenon it indexes: content without traceable support in the seed. Descriptions of how fabrication arises are treated as interpretations. A conceptual typology of fabrication forms distinguishes entity, numerical, relational, causal, epistemic-status and normative forms. These categories were not part of the coder prompt and were not coded. The pipeline also recorded register and specificity ratings, which are not analysed here.

Primary-coder fidelity and completeness scores were heavily heaped. Seventy-one percent of fidelity scores were 97, and 81% of completeness scores were 97 or 99. Their means therefore summarize shifts among a small number of score bins rather than finely resolved interval differences.

Primary-coder parsing. The production pipeline converted each primary-coder reply to a number by removing all non-digit characters and capping the result at 100. Raw replies were not retained. A reply such as "35/100" would therefore be recorded as 100.

Primary-coder hallucination shows the resulting signature: 398 of 5,880 scores are exactly 100, none lies between 36 and 94, and GPT-4.1 and Gemini rated none of the recoded outputs with a primary score of 100 at 90 or above. Their median ratings for those outputs were 35 and 10, respectively. We therefore treat primary-coder hallucination magnitudes as uninterpretable and take hallucination magnitudes from the independent coders, whose replies were parsed strictly.

We therefore use primary-coder hallucination only for rank-based tests, reported with two stress tests: scores of 100 are either set to 35 or treated as missing. These analyses bound the artefact's influence without attempting to recover the original values (Supplementary Section S16).

The same parser was applied to the other primary outcomes. Fidelity and completeness do not show the hallucination ceiling-spike signature, although the missing raw replies prevent a complete assessment of other parser-induced misreadings. Factual retention contains 1,688 ceiling scores, so its primary-coder magnitudes are not interpreted.

Two instruments address epistemic status. A generic three-level item asking whether "the certainty, modality or evidential framing of any claim shifted" was rated by the author audit and two earlier cross-family coders on a 150-pair audit sample (Supplementary Section S5). The proposition-anchored Instrument B is described under Independent recoding.

Units of inference and statistical analysis

For chain-terminal outcomes, the chain is the unit of inference. Recorded outputs within a chain are dependent and are never treated as independent inferential units. Each chain's terminal state is the mean of all of its actual terminal outputs under the common terminal rule. These are Node 11 for the linear chain, Junction C for the ring, the ten spokes for hub-and-spoke, the three lower-tier nodes for the hierarchy, and the fifth post-synthesis node for the lattice and bidirectional lattice.

Primary configuration comparisons use the standard condition, the only condition present for all six architectures, with 35 chains per architecture and 210 chains in total. Pooled analyses across all 12 conditions are treated as sensitivity analyses (Supplementary Section S13).

Omnibus comparisons use Kruskal–Wallis tests with η^2_H together with a seed-stratified permutation test that conditions on the seven tested seeds. The permutation statistic is a van Elteren-type sum, across architectures, of the squared deviation of each architecture's mean within-seed rank from the overall mean. Ranks are scaled within seed, and architecture labels are permuted within seed over 20,000 draws.

Both omnibus procedures treat chains as the units of inference within the seven tested messages. Their p-values therefore describe those messages rather than a population of possible messages. Seed-blocked Friedman tests of architecture means within seed separately assess consistency across the seven messages using the asymptotic χ^2 reference distribution (Supplementary Section S14).

Pairwise comparisons use differences in mean within-seed ranks under the same within-seed permutation, with Holm correction over the 15 architecture pairs for each outcome. Mann–Whitney tests are reported as sensitivity analyses. Because hub-and-spoke and hierarchy terminal states average multiple outputs whereas the other configurations contribute a single terminal, pairwise differences involving these configurations are not interpreted as architecture-specific effects.

Hallucination scores are zero-heavy and right-skewed. For that outcome, we report permutation p-values, Hodges–Lehmann shifts, and means and medians, but not bootstrap confidence intervals.

Cross-lineage divergence

Following the diversification measure used in prior human transmission research [7], cross-lineage divergence (XL-SD) is the standard deviation, across the five lineages in an architecture $\times$ seed cell, of the evolutionary fidelity of the output delivered by each lineage. For configurations with multiple terminals, averaging terminals would mechanically reduce between-lineage dispersion. XL-SD is therefore defined from a single delivered output.

For hub-and-spoke and the hierarchy, each of the $10 \times 3 = 30$ combinations of a delivered spoke and a delivered lower-tier node defines one specification. XL-SD is computed within each specification and then averaged uniformly across all 30 specifications.

The primary test is a blocked randomization test over the seven seed blocks using the Friedman statistic, with architecture labels permuted within seed over 100,000 draws. Asymptotic Friedman and Kruskal–Wallis tests are sensitivity analyses. Rank-order stability and omnibus evidence are reported separately.

Depth-referenced analyses

Four analyses address the entanglement of configuration with transformation depth.

First, a computational-identity placebo compares ordinary single-input transmission nodes that use the same prompt, model and temperature and occupy the same depth in different configurations. A hub spoke and linear Node 2, for example, are both second single-input transformations of the seed. A seed-stratified omnibus test compares the configurations represented at each depth.

Second, the linear chain provides reference means at depths 1–11. Operation contrasts compare each configuration-specific operation with the linear chain at the same depth under both depth definitions. These operations include hierarchy synthesis, ring junctions, lattice and bidirectional synthesis, post-debate states, and downstream nodes following each operation. Because synthesis, junction and debate calls ran at lower temperatures than ordinary transmissions, these contrasts also compare sampling temperatures.

The operation contrasts use seed-stratified permutation tests, with Holm correction within each outcome $\times$ depth-definition family and Bartlett-excluded sensitivity analyses. Nodes deeper than 11 have no linear reference and are not tested.

Finally, bidirectional post-debate XL-SD is compared, paired by seed, with the linear chain at message-carrying-route depth 8 and critical depth 11, with the same bidirectional chains before debate at Chain 5, and with lattice Chain 5, the synthesis input without debate. A pre-debate contrast serves as a placebo. These comparisons use exact two-sided sign-flip tests over the seven seeds.

Independent recoding

Two non-Anthropic coders recoded the outcomes under a protocol deposited before execution that fixed the sample, instruments, model versions, settings, analyses and interpretation rules. This was a prospective revision protocol, not a preregistration of the original simulation study.

The independent coders were OpenAI gpt-4.1-2025-04-14 at temperature 0.1 and Google gemini-3.7-flash at temperature 1.0, the provider default, with thinking level "low". No Anthropic model was used as an independent coder.

The frozen recoding sample contained 2,581 unique outputs, or 44% of all outputs. It comprised all 2,520 outputs required by the terminal and depth-referenced analyses specified in the frozen protocol, plus 61 outputs contributed uniquely by the earlier 150-pair validation sample. Condition labels were hidden. For 517 outputs, each coder completed two additional ratings so that each of those outputs had three ratings for within-coder repeatability.

Instrument A reused the four production coder prompts verbatim. Instrument B used a proposition-anchored codebook containing two source propositions of stated epistemic status, such as a conditional projection, for each organizational seed, giving 12 anchors. For each anchor, coders first judged whether the proposition was present. If present, they classified its status as preserved, strengthened, weakened, altered or uncertain. A shift is defined as strengthened, weakened or altered; uncertain is not counted as a shift. In total, 2,207 outputs were eligible for Instrument B.

All pre-specified analyses, including row-level agreement with Sonnet, repeatability, configuration comparison, divergence, depth-referenced contrasts, capability sensitivity and anchored agreement, were run separately for each coder. Coder results were never pooled.

Interpretation rules were specified in the deposited protocol before execution, with the documented deviations described below and in Supplementary Section S16. A debate result counts as reproduced only if both independent coders match Sonnet's sign and estimate a magnitude between 0.5 and 2.0 times Sonnet's. The protocol required a comparable magnitude, and this numerical band was fixed before any recode results were inspected (Supplementary Section S16, deviation 3). Anchored prevalence statements require between-coder agreement on status shifts of Gwet's AC1 ≥ 0.60 . Measurement disagreement narrows the claim and is never resolved by averaging.

Each coder completed all 17,545 coding jobs with no missing observations. Supplementary Section S16 reports four protocol deviations. The two interpretive deviations clarified underspecified rules before recode results were inspected and withdrew unsupported factual-retention wording after results were produced. No deviation changed an analysis, statistic or threshold.

Capability placement, disclosure and incorporation

Capability placement was evaluated at the chain terminal for hub-and-spoke and the hierarchy by comparing the standard and higher-capability conditions, with 35 chains per condition. Analyses were run both with and without the Bartlett seed using a seed-stratified permutation test of the difference in means, with Mann–Whitney tests as sensitivity analyses.

Disclosure was evaluated by Welch two one-sided tests (TOST) [30] comparing implicit and declared conditions in the lattice and bidirectional lattice. The primary smallest effect size of interest was ± 5 points, with ± 10 points as a sensitivity bound.

Incorporation of the higher-capability source was assessed at each synthesis node using SBERT cosine similarity between the synthesis output and each input chain's final output. A synthesis was classified as A-aligned or B-aligned when the similarity difference was at least 0.05 in the corresponding direction and as blended otherwise. The continuous difference $\Delta(A-B)$ is also reported (Supplementary Section S8).

Robustness

Leave-one-seed-out analyses recompute architecture rankings after excluding each seed in turn. Chain-terminal regression models with seed fixed effects for fidelity and completeness are reported in Supplementary Section S9. Embedding-based similarity, using Sentence-Transformers all-mpnet-base-v2 [31], is used throughout as a distinct semantic instrument and a convergent-validity check rather than as a substitute for the coded outcomes.

Design predictions and analysis transparency

Predictions about architecture-level differences, capability placement, capability disclosure and terminal amplification were specified in the original study design but not preregistered. A timestamped analysis-plan addendum prospectively specified external validation, decomposition, equivalence and incorporation analyses (OSF, DOI 10.17605/OSF.IO/9Z3N4). The regression decomposition is not reported because terminal path length and transformation count are fixed by architecture and its hallucination models used compromised primary-coder magnitudes. The earlier validation also departed from the addendum in coder families and rating items, and the planned primary-coder hallucination equivalence outcome is not interpreted because of the parsing artefact (Supplementary Section S5).

Revision analyses followed the deposited independent-recoding protocol and a separate written revision plan that was not deposited. For the primary coder, the seed-stratified permutation and Holm-corrected pairwise tests, seed-blocked Friedman tests, chain-terminal seed-fixed-effects models, depth-referenced analyses and single-output XL-SD were designed or re-specified after the primary-coder results were known and are therefore post hoc. For the independent coders, analyses fixed in the deposited protocol before their results were inspected are prospective. The seed-blocked Friedman tests and cross-coder depth checks were added later and are post hoc. Instrument B was fixed in the deposited protocol before execution.

Other post hoc additions include the parser diagnostics and stress tests, independent-coder equivalence tests, Bartlett-split summaries, and descriptive inter-coder and factual-retention–

hallucination analyses. Supplementary Sections S1 and S16 provide the full analysis-status map and protocol deviations.

Use of AI in this study

AI models served both as objects of study and as analytical instruments. Claude Haiku 4.5 and Claude Sonnet 4.6 (Anthropic) generated all transmissions, and Sonnet 4.6 also served as the primary coder. GPT-4.1 (OpenAI) and Gemini 3.7 Flash (Google) served as independent coders, while the earlier validation used Gemini 2.5 Flash and GPT-4o-mini. Generative AI tools, including Anthropic Claude and OpenAI ChatGPT, also assisted with editing and with the development and review of analysis code. The authors remain solely accountable for the manuscript, analyses and conclusions.

Data availability

The complete dataset of 5,880 scored outputs from 420 chains, the embedding-similarity outputs and the earlier-validation materials are publicly available through the Open Science Framework (OSF) under DOI 10.17605/OSF.IO/MH9GN [32]. The earlier-validation materials include the blinded audit interface, coding key, author ratings and audit-sample ratings from Gemini 2.5 Flash and GPT-4o-mini.

The same OSF project contains the independent-recoding protocol, frozen sample, instruments and settings deposited before execution. Full recoding outputs from both independent coders, together with the frozen analysis outputs, are available in the project's code and recoding component under DOI 10.17605/OSF.IO/J2ZDU [33]. The analysis-plan addendum for the original validation analyses is available under DOI 10.17605/OSF.IO/9Z3N4.

Code availability

The simulation pipeline used to generate the dataset (simulations/m2m_v2.html, which requires an Anthropic API key), the revision analysis scripts, independent-recoding runner and pre-specified analyses, report generator, and scripted statistical audit are available through the Open Science Framework under DOI 10.17605/OSF.IO/J2ZDU [33].

References

1. Hancock, J. T., Naaman, M. & Levy, K. AI-mediated communication: definition, research agenda, and ethical considerations. *J. Comput.-Mediat. Commun.* **25**, 89–100 (2020).
2. Sundar, S. S. Rise of machine agency: a framework for studying the psychology of human–AI interaction. *J. Comput.-Mediat. Commun.* **25**, 74–88 (2020).
3. Carley, K. A theory of group stability. *Am. Sociol. Rev.* **56**, 331–354 (1991).
4. Centola, D. & Macy, M. Complex contagions and the weakness of long ties. *Am. J. Sociol.* **113**, 702–734 (2007).
5. Rodriguez, M. G., Leskovec, J., Balduzzi, D. & Schölkopf, B. Uncovering the structure and temporal dynamics of information propagation. *Netw. Sci.* **2**, 26–65 (2014).

6. Podolny, J. M. Networks as the pipes and prisms of the market. *Am. J. Sociol.* **107**, 33–60 (2001).
7. Brashears, M. E. & Gladstone, E. Error correction mechanisms in social networks can reduce accuracy and encourage innovation. *Soc. Networks* **44**, 22–35 (2016).
8. Sperber, D. Anthropology and psychology: towards an epidemiology of representations. *Man* **20**, 73–89 (1985).
9. Gladstone, E. *Error in Information Diffusion Processes*. PhD thesis, Cornell Univ. (2016).
10. Mohamed, A., Geng, M., Vazirgiannis, M. & Shang, G. LLM as a broken telephone: iterative generation distorts information. In *Proc. 63rd Annu. Meet. Assoc. Comput. Linguist.* (2025).
11. Perez, J. et al. Analyzing text properties evolution in rephrasing, continuation, and inspiration-taking tasks. In *Proc. 2024 Conf. Empir. Methods Nat. Lang. Process.* (2024).
12. Yax, N., Léger, C., Perez, J. & Oudeyer, P.-Y. When LLMs play the telephone game: cumulative changes and attractors in iterated cultural transmission. In *Proc. 13th Int. Conf. Learn. Represent.* (2025).
13. Kraidia, I., Qaddara, I., Almutairi, A., Alzaben, N. & Belhouari, S. B. When collaboration fails: persuasion driven adversarial influence in multi agent large language model debate. *Sci. Rep.* **16**, 11640 (2026).
14. Wenger, E. & Kenett, Y. N. Large language models are homogeneously creative. *PNAS Nexus* **5**, pgag042 (2026).
15. Papachristou, M. & Yuan, Y. Network formation and dynamics among multi-LLMs. *PNAS Nexus* **4**, pgaf317 (2025).
16. Ashery, A. F., Aiello, L. M. & Baronchelli, A. Emergent social conventions and collective bias in LLM populations. *Sci. Adv.* **11**, eadu9368 (2025).
17. Brinkmann, L. et al. Machine culture. *Nat. Hum. Behav.* **7**, 1855–1868 (2023).
18. Sur, S., Chattopadhyay, R., Claussen, J. C. & Mukhopadhyay, A. A mathematical framework for misinformation propagation in complex networks: topology-dependent distortion and control. Preprint at <https://arxiv.org/abs/2512.19465> (2025).
19. Lin, X. et al. LLM-based agents suffer from hallucinations: a survey of taxonomy, methods, and directions. Preprint at <https://arxiv.org/abs/2509.18970> (2025).
20. Alansari, A. & Luqman, H. Large language models hallucination: a comprehensive survey. Preprint at <https://arxiv.org/abs/2510.06265> (2025).
21. Gagniuc, P. A. Antivirus systems: detection methods and architectures. *Algorithms* **19**, 345 (2026).
22. Bartlett, F. C. *Remembering: A Study in Experimental and Social Psychology* (Cambridge Univ. Press, 1932).
23. Ost, J. et al. The serial reproduction of an urban myth: revisiting Bartlett's schema theory. *Memory* **30**, 1123–1136 (2022).
24. Diaz-Diaz, F., San Miguel, M. & Meloni, S. Echo chambers and information transmission biases in homophilic and heterophilic networks. *Sci. Rep.* **12**, 9350 (2022).
25. Pinheiro, F. L. & Vasconcelos, V. V. Heterogeneous update processes shape information cascades in social networks. *Sci. Rep.* **15**, 13999 (2025).

26. Gagniuc, P. A. Why educated users still click: a review on phishing and the limits of human judgment. *Comput. Sci. Rev.* **62**, 101025 (2026).
27. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
28. Anthropic. Claude Haiku 4.5 model card. <https://www.anthropic.com> (2026).
29. Anthropic. Claude Sonnet 4.6 model card. <https://www.anthropic.com> (2026).
30. Lakens, D. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Soc. Psychol. Personal. Sci.* **8**, 355–362 (2017).
31. Reimers, N. & Gurevych, I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In *Proc. 2019 Conf. Empir. Methods Nat. Lang. Process. Int. Jt. Conf. Nat. Lang. Process. (EMNLP-IJCNLP)* (2019).
32. Gladstone, E. & Brashears, M. E. Information mutation in AI-agent communication networks. *Open Science Framework* <https://doi.org/10.17605/OSF.IO/MH9GN> (2026).
33. Gladstone, E. & Brashears, M. E. Code and recoding release for "Information mutation across implemented AI-agent communication architectures". *Open Science Framework* <https://doi.org/10.17605/OSF.IO/J2ZDU> (2026).

Funding

The authors received no specific funding for this work.

Author contributions

E.G. conceived and designed the study, implemented the simulation and analysis software, performed the validation and formal analysis, curated the data, produced the visualizations and wrote the original draft. M.E.B. contributed to the conceptualization and methodology and provided supervision. Both authors reviewed, edited and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Supplementary Materials

Manuscript: Information mutation across implemented AI-agent communication architectures

Authors: Gladstone & Brashears

Data and protocols: OSF DOI 10.17605/OSF.IO/MH9GN, analysis-plan addendum OSF DOI 10.17605/OSF.IO/9Z3N4, code and recoding outputs OSF DOI 10.17605/OSF.IO/J2ZDU

All values in these materials are computed by the deposited analysis scripts, and recoding values come from the frozen recoding outputs (Section S16). Unless stated otherwise, analyses use the standard condition, the chain as the unit and the common terminal rule (Methods).

Section S1. Original design predictions and their current status

The original design specified seven predictions. They were specified in the design but not preregistered. The table maps each to its current evidentiary status.

Supplementary Table S1. Original design predictions and their current evidentiary status.

Original prediction	Current evidentiary status	Where reported
H1: Architecture affects fabrication through synthesis at multi-input nodes	Not tested as a mechanism. Terminal fabrication ordering was directionally similar under the independent coders but less consistent across messages and coders than fidelity, and the cross-configuration ordering cannot be separated from transformation depth.	Main text: "Terminal mutation profiles", "Transformation depth"
H2: Centralized architectures produce lower cross-lineage divergence than synthesis architectures	Detected under the primary coder only (post hoc estimand); not reproduced by either independent coder or the embedding measure. No general claim is made.	Main text: "Cross-lineage divergence"; S12
H3: Higher-capability models at structurally amplified positions reduce fabrication	Not supported at the chain terminal; the primary coder's contrast is unstable under parser stress tests; independent-coder evidence mixed	Main text: "Capability placement"; S16 (capability placement by coder)
H4: Declared status does not improve synthesis relative to implicit capability asymmetry	Consistent: no improvement detected; equivalence within ± 5 for fidelity and completeness; bidirectional hallucination not equivalent under either independent coder	S7, S8
H5: Architecture effects on fabrication are larger for narratively complex (Bartlett) content	Exploratory; one stimulus	Main text: "Fabrication and the Bartlett stress test"; S11
H6: Architecture effects are larger at terminal outputs than at intermediate positions	Replaced by the depth-referenced analysis. Node positions are not comparable across architectures (S6).	S6
H7: Capability condition does not reduce cross-lineage divergence after accounting for architecture	No dedicated test in the final analysis	S12, S13

Several analyses were specified in a timestamped addendum deposited before execution: external validation (S5), decomposition (not reported; see S6 for depth-referenced analyses), equivalence testing (S7) and incorporation (S8). Section S5 reports departures from that addendum in the earlier validation. Analyses added during revision drew on a written revision analysis plan that

was not deposited and, for the independent recoding, a protocol deposited before execution (S16). Analyses added after the frozen recoding protocol are labelled post hoc where reported.

Section S2. Capability framework

We distinguish the *placement* of a higher-capability model from its *disclosure* to downstream agents. These are design labels and imply no status-like cognition in the agents.

- **Placement.** Claude Sonnet 4.6 occupies the hub (hub-and-spoke) or the upper tier (directed hierarchy), whose outputs reach many downstream nodes.
- **Asymmetry.** In the lattice and bidirectional lattice, Sonnet 4.6 produces the five transmissions of Chain A while Chain B uses the baseline model. In the bidirectional lattice, Chain A's output is then revised R times by the baseline model during debate before synthesis (default 3, not recorded per chain).
- **Disclosure.** In the declared condition, the synthesis node is told that Version A was processed by a higher-capability system. The implicit condition has the same asymmetry without the note.
- **Incorporation.** The behavioural outcome of disclosure is the preferential incorporation of the labelled source, measured in S8.

Section S3. Architecture taxonomy and human stages

The six architectures sit within a broader family of automated LLM-agent communication structures:

- **sequential:** the linear chain, the ring with shortcuts, a small-world variant
- **centralized and tiered:** hub-and-spoke, the directed hierarchy
- **parallel-chain synthesis:** the lattice, the bidirectional lattice, analogous to multi-agent debate or cross-review.

Hybrid structures with a human at the terminal or at an intermediate checkpoint are outside the present scope. A human stage should not be treated as a corrective with ground-truth status. Human correction in transmission networks can reduce accuracy (main-text ref. 7), and human judgment of authoritative-seeming information under uncertainty is itself fallible (main-text ref. 26). A human checkpoint should therefore not be assumed to restore ground truth or eliminate upstream distortion.

Section S4. Prompts

The prompts below are reproduced verbatim from the deposited simulation code (simulations/m2m_v2.html), which the development log records as the production file. Prompts were not stored per call, so the attribution rests on the development log, the deposited file and the output-length check below. $\{\dots\}$ marks runtime substitutions.

Model settings.

- Temperatures: transmission 0.70, synthesis and ring junctions 0.30, debate feedback 0.30, debate revision 0.50, primary coder 0.10.
- Maximum 500 output tokens per call.
- The production run recorded one coder rating per outcome per output.

Outputs were not progressively shortened. Mean output length relative to the seed was 1.10 across all 5,880 recorded outputs, 1.08 at the first linear node and 1.09 at the eleventh (words; standard-condition linear chains for the node-specific values).

Transmission node (active mediation, all analysed runs).

You are processing a message in an information network. Read the following message and rewrite it to accurately capture its meaning. $\text{\texttt{\$ \{fmtInstr(format)\}}}$

Message: $\text{\texttt{\$ \{text\}}}$

Respond with only your rewritten version, nothing else.

Synthesis node (lattice and bidirectional synthesis, hierarchy mid tier, ring junctions).

You are a synthesis node in an information network. You have received two versions of a message from parallel channels. Produce a single unified message that best captures the accurate combined meaning. $\text{\texttt{\$ \{authorityNote ? \}}}$

$\text{\texttt{' +authorityNote : " \} \{fmtInstr(format)\}}}$

Version A: $\text{\texttt{\$ \{textA\}}}$

Version B: $\text{\texttt{\$ \{textB\}}}$

Respond with only your synthesized message, nothing else.

Debate feedback (bidirectional lattice, showing the original seed).

You are reviewing a message that was derived from an original. Provide one sentence of constructive feedback on what meaning may have been lost or changed.

Derived message: $\text{\texttt{\$ \{text\}}}$

Original: $\text{\texttt{\$ \{original\}}}$

Respond with only your one-sentence feedback.

Debate revision (bidirectional lattice).

You are refining a message based on feedback from a peer node. Revise your draft to address the feedback while preserving accuracy. $\text{\texttt{\$ \{fmtInstr(format)\}}}$

Your draft: "\${draft}"
Peer feedback: "\${feedback}"

Respond with only your revised message, nothing else.

Passive relay (calibration only, not analysed).

You are a relay node in an information network. Reproduce the following message as accurately as possible. \${fmtInstr(format)}

Message: "\${text}"

Respond with only the message, nothing else.

Coder: evolutionary fidelity (the two texts are the seed and the scored output).

Rate the semantic similarity of these two texts from 0 to 100.
0 = completely different meanings. 100 = identical meanings.
Focus only on meaning — ignore spelling, grammar, and format differences.

Text A: "\${a}"
Text B: "\${b}"

Respond with only a single integer.

Coder: completeness.

Rate information completeness: what proportion of the semantic content from the original is preserved in the transmission?
0 = nothing preserved. 100 = all content preserved.
Focus on content coverage, not wording.

Original: "\${seed}"
Transmission: "\${transmission}"

Respond with only a single integer from 0 to 100.

Coder: factual retention.

You are assessing factual retention. The original seed contains specific factual details (numbers, names, dates, percentages, thresholds). Rate what proportion of those specific factual details survive intact in the transmission, from 0 to 100.
0 = no factual details preserved. 100 = all factual details preserved exactly.

Original seed: "\${seed}"
Transmission: "\${transmission}"

Respond with only a single integer from 0 to 100.

Coder: hallucination.

You are assessing hallucination in a message transmission. Rate the degree to which the transmission contains specific details, claims, or facts that are NOT present in the original, from 0 to 100.

0 = no hallucinated content (everything stated in the transmission can be found in the original). 100 = severely hallucinated (transmission contains many fabricated details absent from the original).

Focus on added factual content — numbers, names, dates, causal claims, events — that cannot be found in the original. Paraphrase and synonym substitution alone do not count as hallucination.

Original: "\${seed}"

Transmission: "\${transmission}"

Respond with only a single integer from 0 to 100.

`$(fmtInstr(format))` resolves to: Use standard written English. In the declared condition, authorityNote is: Note: Version A was processed by a higher-capability system.

The prompts of the independent recoding are described in S16: Instrument A reuses the four coder prompts above verbatim.

Section S5. Earlier validation (descriptive)

The coded outcomes were also evaluated against:

- the embedding similarity of every output to its seed (Sentence-Transformers all-mpnet-base-v2)
- a blinded audit by one author (blind to architecture, condition, seed, position and model scores) of a stratified sample of transmission pairs
- two cross-family coders, Gemini 2.5 Flash and GPT-4o-mini, on the same audit sample.

A stratified sample of 150 transmission pairs was drawn. One author completed blinded ratings for 149 pairs; one pair has no saved author rating. Gemini 2.5 Flash returned ratings for 149 of 150 pairs because one request failed after retries, and GPT-4o-mini rated all 150. The generic epistemic-status agreement analyses use the 148 pairs with complete ratings from all three raters.

The earlier validation departed from the timestamped addendum (OSF DOI 10.17605/OSF.IO/9Z3N4) in several respects: Gemini 2.5 Flash was used instead of Gemini 2.5 Pro, GPT-4o-mini was added as a second cross-family coder, and the cross-family ratings used the audit's ordinal added-content and three-level epistemic-status items rather than the planned 0–100 hallucination and binary-status items. The addendum designated primary-coder

hallucination as the primary equivalence outcome; that result is not interpreted because the production hallucination scale is affected by the parsing artefact, and the independent-coder equivalence analyses (Section S7) are post hoc and are not presented as fulfilling the addendum.

These checks are descriptive. The audit and the two cross-family coders covered too few outputs per architecture to support architecture-level comparisons, and the independent recoding of the complete terminal and depth-referenced sample (S16) carries the validation weight.

Supplementary Table S2. *Earlier validation: per-output agreement with the primary coder (Spearman ρ ; $n = 149$ author, 149 Gemini 2.5 Flash, 150 GPT-4o-mini).*

Source	Similarity vs fidelity	Added content vs hallucination
Author audit	.51	.56
Gemini 2.5 Flash	.34	.52
GPT-4o-mini	.67	.55

Embedding similarity (standard condition): the embedding measure and the primary coder's terminal fidelity gave:

- the same architecture ordering ($\rho = 1.00$)
- a chain-level correlation of $\rho = .54$.

Generic epistemic-status item. The audit and both earlier cross-family coders also answered a generic item: "Has the certainty, modality, or evidential framing of any claim shifted?" (no / yes / uncertain). Agreement on "yes" was poor, so the item is not used to estimate prevalence. The audit sample contained 150 pairs, of which 148 had complete generic ratings from all three raters; agreement statistics use those 148.

Supplementary Table S3. Generic epistemic-status item: agreement (150 pairs sampled; 148 complete across all three raters).

Pair	Raw agreement	Cohen κ	Gwet AC1
Audit vs Gemini 2.5 Flash	63%	.29	.26
Audit vs GPT-4o-mini	47%	.04	.04
Gemini 2.5 Flash vs GPT-4o-mini	68%	.15	.51

"Yes" rates were 57% (audit), 33% (Gemini 2.5 Flash) and 12% (GPT-4o-mini). Fleiss' $\kappa = .10$.

Section S6. Transformation depth and depth-referenced analyses

Four depth quantities

Stored positions in the dataset are not comparable across architectures. The table below gives, for each architecture:

- recorded outputs and model calls per chain
- the stored positions of the terminal outputs
- the two depth definitions (Methods).

Bidirectional values assume $R = 3$ debate rounds.

Supplementary Table S4. *Recorded outputs, model calls, stored terminal positions and the two depth definitions, by architecture.*

Architecture	Recorded outputs per chain	Model calls per chain	Stored terminal position(s)	Shortest message-carrying route	Critical depth
Linear chain	11	11	11	11	11
Ring with shortcuts	11	14	11	6	11
Hub-and-spoke	11	11	2, 3, 4, 5, 6, 7, 8, 9, 10, 11	2	2
Directed hierarchy	11	11	9, 10, 11	3	3
Lattice	16	16	11	11	11
Bidirectional lattice	18	28	11	14	17

Note: Hub-and-spoke spokes are stored at positions 2–11 but are all depth 2. Ring shortcut transmissions are model calls that are not stored as rows, and bidirectional feedback calls and all but the final revision of each chain are not stored as rows. The message-carrying route counts only transformations that carry the message itself. Debate feedback calls also receive the original seed as a reference input. Counted as a route, that re-exposure would place the bidirectional post-debate states at depth 2, its synthesis node at 3 and its terminal output at 8. The depth contrasts are not recomputed under that alternative.

Computational-identity placebo

At each depth, ordinary single-input transmission nodes (same prompt, model and temperature) are compared across the architectures that have such a node at that depth: all six at depth 1; the linear chain, ring, hub-and-spoke (spokes), lattice and bidirectional lattice at depth 2; the linear chain, ring, lattice and bidirectional lattice at depth 3; and the linear chain, lattice and bidirectional lattice at depths 4 and 5. The table gives the seed-stratified omnibus permutation p across those architectures. It is not a pairwise hub-versus-linear test.

Supplementary Table S5. *Computational-identity placebo: seed-stratified omnibus permutation p across the architectures present at each depth.*

Depth	Fidelity p	Completeness p
1	.511	.745
2	.392	.356
3	.807	.042
4	.763	.505
5	.866	.306

Linear reference curve

Standard-condition linear-chain means by depth (primary coder):

Supplementary Table S6. *Linear-chain reference means by depth (primary coder, standard condition). Hallucination by depth under the independent coders is in the cross-coder table below.*

Depth	Fidelity	Completeness
1	96.97	98.43
2	96.74	97.97
3	96.20	98.17
4	96.17	97.86
5	95.54	97.60
6	95.54	97.80
7	95.43	97.46
8	94.71	97.37
9	94.66	97.14
10	94.91	97.03
11	94.83	96.91

Cross-coder depth checks (post hoc)

These checks were added after the independent recoding was complete, so they are not pre-specified. They apply the definitions used above for the primary coder to the frozen outputs of the two independent coders. The recoding sample covered every node required by the protocol's analyses, including every linear-chain node. However, it contained almost none of the ring, lattice and bidirectional transmission nodes at depths 1 to 4. The five-architecture identity test at depth 2 therefore cannot be repeated for the independent coders. Instead, hub spokes are compared with linear Node 2, the comparison on which the shallow-depth argument rests, for all three coders. The depth-5 test uses the same architectures as the primary coder's test.

Supplementary Table S7. *Linear-chain means by depth under the independent coders (standard condition, 35 chains per depth).*

Depth	Hallucination: GPT-4.1	Hallucination: Gemini	Fidelity: GPT-4.1	Fidelity: Gemini	Completeness: GPT-4.1	Completeness: Gemini
1	3.4	0.6	97.1	98.5	97.8	100.0
2	8.0	2.4	96.7	97.9	97.4	100.0
3	8.1	2.4	96.6	97.9	97.4	99.9
4	9.4	3.4	96.2	97.7	97.2	100.0
5	10.1	3.7	96.1	97.6	96.6	100.0
6	10.3	3.9	96.0	97.4	96.7	100.0
7	10.7	4.1	96.1	97.2	96.7	99.9
8	11.6	5.2	95.8	97.1	96.2	99.9
9	11.9	4.7	95.9	97.1	96.3	99.7
10	11.1	5.1	95.8	97.3	96.0	99.9
11	11.0	4.1	95.6	97.3	96.3	99.6

Supplementary Table S8. *Identity checks by coder: seed-stratified permutation p, with hallucination means in parentheses (independent coders).*

Coder	Depth 2, hub spokes vs linear Node 2: hallucination	Fidelity	Completeness	Depth 5, linear, lattice and bidirectional: hallucination	Fidelity	Completeness
Sonnet 4.6 (primary)	—	.319	.354	—	.866	.306
GPT-4.1	.919 (hub 7.1, linear 8.0)	.250	.196	.688	.818	.888

Coder	Depth 2, hub spokes vs linear Node 2: hallucination	Fidelity	Completeness	Depth 5, linear, lattice and bidirectional: hallucination	Fidelity	Completeness
Gemini 3.7 Flash	.081 (hub 1.4, linear 2.4)	.398	.018	.658	.922	.019

Gemini's two completeness differences lie at the ceiling of the scale. The linear means are 100.00 at depth 2 and 100.00 at depth 5, against 99.93 for hub spokes and 99.84 and 99.60 for the lattice and bidirectional chains. Identical computations (same prompt, model and temperature) need not produce identical outcome distributions, and these checks are not equivalence tests. This post hoc check compares each chain's mean across ten hub spokes with a single linear Node 2 output, the averaged-versus-single comparison noted in Methods; it is not an equivalence test.

Operation contrasts at matched depth, by coder

Each configuration-specific operation is compared with the linear chain at the same depth. Differences are operation minus linear, and Holm p-values are computed within each outcome $\times$ definition family. Because synthesis, junction and debate calls ran at lower temperatures than ordinary transmissions, these operation contrasts also compare sampling temperatures. Hallucination contrasts are shown for the independent coders only (Methods, primary-coder parsing). Nodes deeper than 11 have no linear reference and are not tested. Bartlett-excluded differences are in the OSF outputs.

Supplementary Table S9. *Hallucination, critical depth: operation minus linear-chain mean at the same depth, by coder (Holm p within the outcome $\times$ definition family). Independent coders only (Methods, primary-coder parsing).*

Operation	Depth	GPT-4.1 diff (Holm p)	Gemini diff (Holm p)
hierarchy mid-tier synthesis	2	-4.46 (.002)	-2.26 (.008)
hierarchy lower tier	3	-0.48 (> .99)	-0.50 (> .99)
ring Junction A	4	-1.14 (> .99)	-2.17 (.702)
ring nodes after Junction A	5	-0.43 (> .99)	-1.57 (> .99)
lattice synthesis	6	-0.29 (> .99)	-0.66 (> .99)
ring nodes after Junction A	6	-0.86 (> .99)	-1.80 (> .99)
lattice post-synthesis	7	-0.86 (> .99)	+0.09 (> .99)
ring nodes after Junction A	7	+0.00 (> .99)	-1.40 (> .99)
lattice post-synthesis	8	-0.43 (> .99)	-0.77 (> .99)
ring Junction B	8	-1.14 (> .99)	-2.63 (.906)
lattice post-synthesis	9	-2.43 (> .99)	-0.57 (> .99)
ring nodes after Junction B	9	-0.71 (> .99)	-0.63 (> .99)
lattice post-synthesis	10	+0.29 (> .99)	-1.14 (> .99)
ring nodes after Junction B	10	+1.43 (> .99)	-1.86 (> .99)
bidirectional post-debate states	11	+0.14 (> .99)	+1.00 (> .99)
lattice post-synthesis	11	+0.29 (> .99)	+0.00 (> .99)
ring Junction C (terminal)	11	+0.00 (> .99)	-1.20 (> .99)

Supplementary Table S10. *Hallucination, shortest message-carrying route: operation minus linear-chain mean at the same depth, by coder (Holm p within the outcome $\times$ definition family). Independent coders only (Methods, primary-coder parsing).*

Operation	Depth	GPT-4.1 diff (Holm p)	Gemini diff (Holm p)
hierarchy mid-tier synthesis	2	-4.46 (.002)	-2.26 (.010)
ring Junction A	2	+0.29 (> .99)	-1.20 (> .99)

Operation	Depth	GPT-4.1 diff (Holm p)	Gemini diff (Holm p)
hierarchy lower tier	3	-0.48 (> .99)	-0.50 (> .99)
ring nodes after Junction A	3	+1.57 (> .99)	-0.29 (> .99)
ring Junction B	4	+1.00 (> .99)	-0.83 (> .99)
ring nodes after Junction A	4	+0.00 (> .99)	-1.26 (> .99)
ring nodes after Junction A	5	+0.57 (> .99)	-1.06 (> .99)
ring nodes after Junction B	5	+1.00 (> .99)	+0.37 (> .99)
lattice synthesis	6	-0.29 (> .99)	-0.66 (> .99)
ring Junction C (terminal)	6	+0.71 (> .99)	-1.00 (> .99)
ring nodes after Junction B	6	+2.29 (> .99)	-0.66 (> .99)
lattice post-synthesis	7	-0.86 (> .99)	+0.09 (> .99)
bidirectional post-debate states	8	-0.43 (> .99)	-0.06 (> .99)
lattice post-synthesis	8	-0.43 (> .99)	-0.77 (> .99)
bidirectional synthesis	9	+1.71 (> .99)	+2.63 (> .99)
lattice post-synthesis	9	-2.43 (> .99)	-0.57 (> .99)
bidirectional post-synthesis	10	+5.86 (.089)	+2.86 (> .99)
lattice post-synthesis	10	+0.29 (> .99)	-1.14 (> .99)
bidirectional post-synthesis	11	+3.71 (.954)	+4.86 (.049)
lattice post-synthesis	11	+0.29 (> .99)	+0.00 (> .99)

Supplementary Table S11. *Evolutionary fidelity, critical depth: operation minus linear-chain mean at the same depth, by coder (Holm p within the outcome × definition family).*

Operation	Depth	Sonnet diff (Holm p)	GPT-4.1 diff (Holm p)	Gemini diff (Holm p)
hierarchy mid-tier synthesis	2	+0.23 (.235)	+0.18 (> .99)	+0.42 (.669)
hierarchy lower tier	3	+0.34 (> .99)	-0.11 (> .99)	-0.20 (> .99)
ring Junction A	4	+0.69 (.235)	+0.40 (> .99)	+0.20 (> .99)
ring nodes after Junction A	5	+1.09 (.034)	+0.31 (> .99)	+0.00 (> .99)
lattice synthesis	6	+0.43 (> .99)	+0.00 (> .99)	+0.11 (> .99)
ring nodes after Junction A	6	+1.09 (.045)	+0.57 (> .99)	+0.40 (> .99)
lattice post-synthesis	7	+0.06 (> .99)	-0.29 (> .99)	+0.06 (> .99)
ring nodes after Junction A	7	+0.31 (> .99)	+0.14 (> .99)	+0.34 (> .99)
lattice post-synthesis	8	+0.46 (> .99)	-0.20 (> .99)	+0.11 (> .99)
ring Junction B	8	+1.40 (.185)	+0.23 (> .99)	+0.26 (> .99)
lattice post-synthesis	9	+0.63 (> .99)	+0.03 (> .99)	+0.29 (> .99)
ring nodes after Junction B	9	+1.54 (.072)	+0.20 (> .99)	+0.26 (> .99)
lattice post-synthesis	10	+0.37 (> .99)	+0.17 (> .99)	-0.23 (> .99)
ring nodes after Junction B	10	+0.77 (> .99)	-0.17 (> .99)	+0.06 (> .99)
bidirectional post-debate states	11	-0.84 (> .99)	-0.11 (> .99)	-1.56 (.027)
lattice post-synthesis	11	+0.43 (> .99)	-0.11 (> .99)	-0.23 (> .99)
ring Junction C (terminal)	11	+1.03 (> .99)	+0.46 (> .99)	+0.06 (> .99)

Supplementary Table S12. *Evolutionary fidelity, shortest message-carrying route: operation minus linear-chain mean at the same depth, by coder (Holm p within the outcome × definition family).*

Operation	Depth	Sonnet diff (Holm p)	GPT-4.1 diff (Holm p)	Gemini diff (Holm p)
hierarchy mid-tier synthesis	2	+0.23 (.328)	+0.18 (> .99)	+0.42 (.669)
ring Junction A	2	+0.11 (> .99)	-0.09 (> .99)	-0.09 (> .99)
hierarchy lower tier	3	+0.34 (> .99)	-0.11 (> .99)	-0.20 (> .99)
ring nodes after Junction A	3	+0.43 (> .99)	-0.23 (> .99)	-0.34 (> .99)
ring Junction B	4	-0.06 (> .99)	-0.23 (> .99)	-0.26 (> .99)
ring nodes after Junction A	4	+0.46 (> .99)	+0.31 (> .99)	+0.11 (> .99)
ring nodes after Junction A	5	+0.20 (> .99)	+0.17 (> .99)	-0.06 (> .99)
ring nodes after Junction B	5	+0.66 (> .99)	+0.03 (> .99)	-0.26 (> .99)
lattice synthesis	6	+0.43 (> .99)	+0.00 (> .99)	+0.11 (> .99)
ring Junction C (terminal)	6	+0.31 (> .99)	+0.11 (> .99)	-0.06 (> .99)
ring nodes after Junction B	6	+0.14 (> .99)	-0.31 (> .99)	-0.06 (> .99)
lattice post-synthesis	7	+0.06 (> .99)	-0.29 (> .99)	+0.06 (> .99)
bidirectional post-debate states	8	-0.73 (> .99)	-0.26 (> .99)	-1.44 (.085)
lattice post-synthesis	8	+0.46 (> .99)	-0.20 (> .99)	+0.11 (> .99)

Operation	Depth	Sonnet diff (Holm p)	GPT-4.1 diff (Holm p)	Gemini diff (Holm p)
bidirectional synthesis	9	-1.03 (> .99)	-0.29 (> .99)	-1.80 (.052)
lattice post-synthesis	9	+0.63 (> .99)	+0.03 (> .99)	+0.29 (> .99)
bidirectional post-synthesis	10	-2.37 (.084)	-0.80 (.520)	-2.26 (.002)
lattice post-synthesis	10	+0.37 (> .99)	+0.17 (> .99)	-0.23 (> .99)
bidirectional post-synthesis	11	-2.06 (.118)	-0.63 (> .99)	-2.34 (.002)
lattice post-synthesis	11	+0.43 (> .99)	-0.11 (> .99)	-0.23 (> .99)

Supplementary Table S13. *Completeness, critical depth: operation minus linear-chain mean at the same depth, by coder (Holm p within the outcome × definition family).*

Operation	Depth	Sonnet diff (Holm p)	GPT-4.1 diff (Holm p)	Gemini diff (Holm p)
hierarchy mid-tier synthesis	2	+0.39 (.434)	+0.26 (> .99)	-0.07 (> .99)
hierarchy lower tier	3	-0.09 (> .99)	-0.37 (> .99)	-0.07 (> .99)
ring Junction A	4	+0.09 (> .99)	-0.20 (> .99)	+0.00 (> .99)
ring nodes after Junction A	5	+0.17 (> .99)	-0.34 (> .99)	-0.06 (> .99)
lattice synthesis	6	-0.46 (> .99)	-0.46 (> .99)	-0.26 (> .99)
ring nodes after Junction A	6	-0.03 (> .99)	+0.17 (> .99)	-0.06 (> .99)
lattice post-synthesis	7	-0.91 (> .99)	-0.77 (> .99)	-0.23 (> .99)
ring nodes after Junction A	7	-0.29 (> .99)	-0.71 (> .99)	-0.23 (> .99)
lattice post-synthesis	8	-1.00 (> .99)	-0.40 (> .99)	-0.49 (> .99)
ring Junction B	8	+0.03 (> .99)	+0.63 (> .99)	+0.00 (> .99)
lattice post-synthesis	9	-0.40 (> .99)	-0.49 (> .99)	-0.37 (> .99)
ring nodes after Junction B	9	+0.20 (> .99)	+0.20 (> .99)	+0.06 (> .99)
lattice post-synthesis	10	-1.14 (> .99)	-0.43 (> .99)	-0.40 (> .99)
ring nodes after Junction B	10	+0.23 (> .99)	-0.06 (> .99)	+0.00 (> .99)
bidirectional post-debate states	11	-1.99 (.088)	-1.14 (> .99)	-1.54 (.037)
lattice post-synthesis	11	-0.51 (> .99)	-0.49 (> .99)	-0.11 (> .99)
ring Junction C (terminal)	11	+0.49 (> .99)	-0.23 (> .99)	+0.29 (> .99)

Supplementary Table S14. *Completeness, shortest message-carrying route: operation minus linear-chain mean at the same depth, by coder (Holm p within the outcome × definition family).*

Operation	Depth	Sonnet diff (Holm p)	GPT-4.1 diff (Holm p)	Gemini diff (Holm p)
hierarchy mid-tier synthesis	2	+0.39 (.434)	+0.26 (> .99)	-0.07 (> .99)
ring Junction A	2	-0.03 (> .99)	-0.43 (> .99)	+0.00 (> .99)
hierarchy lower tier	3	-0.09 (> .99)	-0.37 (> .99)	-0.07 (> .99)
ring nodes after Junction A	3	-0.40 (.759)	-1.11 (.176)	+0.00 (> .99)
ring Junction B	4	-0.46 (.816)	-0.34 (> .99)	-0.14 (> .99)
ring nodes after Junction A	4	-0.09 (> .99)	-0.34 (> .99)	-0.06 (> .99)
ring nodes after Junction A	5	-0.43 (> .99)	-0.63 (> .99)	-0.34 (> .99)
ring nodes after Junction B	5	-0.26 (> .99)	-0.06 (> .99)	-0.20 (> .99)
lattice synthesis	6	-0.46 (> .99)	-0.46 (> .99)	-0.26 (> .99)
ring Junction C (terminal)	6	-0.40 (> .99)	-0.57 (> .99)	-0.14 (> .99)
ring nodes after Junction B	6	-0.54 (.759)	-0.74 (> .99)	-0.14 (> .99)
lattice post-synthesis	7	-0.91 (> .99)	-0.77 (> .99)	-0.23 (> .99)
bidirectional post-debate states	8	-2.44 (.009)	-1.03 (> .99)	-1.83 (.002)
lattice post-synthesis	8	-1.00 (> .99)	-0.40 (> .99)	-0.49 (> .99)
bidirectional synthesis	9	-1.69 (.389)	-0.97 (> .99)	-0.91 (> .99)
lattice post-synthesis	9	-0.40 (> .99)	-0.49 (> .99)	-0.37 (> .99)
bidirectional post-synthesis	10	-2.26 (.004)	-1.00 (.410)	-1.11 (.160)
lattice post-synthesis	10	-1.14 (> .99)	-0.43 (> .99)	-0.40 (> .99)
bidirectional post-synthesis	11	-1.77 (.027)	-1.11 (.709)	-1.09 (.684)
lattice post-synthesis	11	-0.51 (> .99)	-0.49 (> .99)	-0.11 (> .99)

These tables contain the full contrast family specified in the independent-recoding protocol. For the primary coder, the depth-referenced analyses were designed during revision after the primary-coder results were known and are therefore post hoc; their application to the

independent coders was fixed before those coders' results were inspected. The main text discusses:

- the hierarchy mid-tier fabrication contrast
- the Gemini-only bidirectional post-synthesis fabrication contrast
- the post-debate completeness contrasts.

The remaining Holm-surviving contrasts are also coder- and definition-specific:

- ring nodes after Junction A (fidelity, primary coder, critical definition)
- bidirectional post-synthesis nodes (completeness under the primary coder, message-carrying route; fidelity under Gemini, message-carrying route)
- bidirectional post-debate states (fidelity under Gemini, critical definition).

They are reported here and not interpreted further.

Debate and divergence, by coder

Mean seed-level paired difference in XL-SD (contrast minus reference), with the number of seeds with a positive difference and the exact two-sided sign-flip p over seven seeds:

Supplementary Table S15. *Debate-associated change in cross-lineage divergence, by coder and embedding measure.*

Contrast	Sonnet 4.6	GPT-4.1	Gemini 3.7 Flash	SBERT
Post-debate vs linear, message-carrying route depth 8	+1.45 (7/7; p .016)	+0.59 (5/7; p .125)	+1.67 (7/7; p .016)	−0.11 (p .891)
Post-debate vs linear, critical depth 11	+1.44 (6/7; p .047)	+0.67 (4/7; p .219)	+1.86 (7/7; p .016)	−0.41 (p .578)
Post-debate vs own pre-debate chains	+1.62 (7/7; p .016)	+0.86 (5/7; p .062)	+1.71 (7/7; p .016)	+0.18 (p .656)
Placebo: pre-debate vs linear depth 5	+0.54 (4/7; p .375)	−0.251 (2/7; p .469)	−0.002 (4/7; p > .99)	−0.48 (p .766)

The reproduction rule for a headline claim required both independent coders to show the same sign and 0.5–2.0 times the primary coder's magnitude (the protocol required a comparable magnitude, and this band was fixed before any recode results were inspected; Section S16, deviation 3) on the depth-8 and within-architecture contrasts. The rule was not met because the GPT-4.1 depth-8 magnitude fell outside the range.

Section S7. Equivalence tests for capability disclosure

Welch two one-sided tests (TOST) of implicit versus declared conditions at the chain terminal (35 chains each). The primary smallest effect size of interest is ± 5 points. ± 10 is a sensitivity bound. Intervals are 90% confidence intervals for implicit minus declared.

Supplementary Table S16. *Equivalence tests for capability disclosure, primary coder (implicit minus declared).*

Architecture	Outcome	90% CI	Equivalent at ± 5	Equivalent at ± 10
Lattice	Evolutionary fidelity	−2.53 to +1.61	1	1
Lattice	Embedding fidelity	−1.58 to +1.91	1	1
Lattice	Completeness	−1.27 to +1.84	1	1
Bidirectional	Evolutionary fidelity	−1.28 to +3.40	1	1

Architecture	Outcome	90% CI	Equivalent at ± 5	Equivalent at ± 10
Bidirectional	Embedding fidelity	-1.15 to +2.52	1	1
Bidirectional	Completeness	-1.00 to +3.06	1	1

Note: 1 = equivalent, 0 = not equivalent.

The primary-coder table omits hallucination and factual retention, whose primary-coder magnitudes are affected by the parsing artefact (Methods; Section S16). The independent coders, below, carry the disclosure evidence for those outcomes.

Supplementary Table S17. *Equivalence tests for capability disclosure under the independent coders (implicit minus declared; post hoc).*

Coder	Architecture	Outcome	90% CI	Equivalent at ± 5	Equivalent at ± 10
GPT-4.1	Lattice	Hallucination	-4.74 to +7.60	0	1
GPT-4.1	Lattice	Evolutionary fidelity	-1.00 to +0.77	1	1
GPT-4.1	Lattice	Completeness	-1.91 to +1.28	1	1
GPT-4.1	Lattice	Factual retention	-3.35 to +3.92	1	1
GPT-4.1	Bidirectional	Hallucination	-8.38 to +3.53	0	1
GPT-4.1	Bidirectional	Evolutionary fidelity	-0.49 to +1.40	1	1
GPT-4.1	Bidirectional	Completeness	-0.86 to +2.00	1	1
GPT-4.1	Bidirectional	Factual retention	-1.26 to +3.83	1	1
Gemini	Lattice	Hallucination	-3.61 to +2.58	1	1
Gemini	Lattice	Evolutionary fidelity	-1.36 to +0.56	1	1
Gemini	Lattice	Completeness	-0.91 to +0.68	1	1
Gemini	Lattice	Factual retention	-1.56 to +1.68	1	1
Gemini	Bidirectional	Hallucination	-6.94 to +2.37	0	1
Gemini	Bidirectional	Evolutionary fidelity	-0.63 to +2.29	1	1
Gemini	Bidirectional	Completeness	-0.74 to +1.83	1	1
Gemini	Bidirectional	Factual retention	-1.52 to +1.29	1	1

Note: 1 = equivalent, 0 = not equivalent. Each coder has 35 chains per condition, with no missing terminal scores. Under both coders, fidelity, completeness and factual retention are equivalent at ± 5 in both architectures, and the primary coder's bidirectional factual-retention difference is not reproduced. Bidirectional hallucination is not equivalent at ± 5 under either coder.

Section S8. Incorporation of the higher-capability source

At each lattice and bidirectional synthesis node, SBERT cosine similarity was computed between the synthesis output and the final output of each input chain. A decision was classified as A-aligned if similarity to Chain A exceeded similarity to Chain B by at least 0.05, B-aligned under the reverse criterion, and blended otherwise.

A-aligned decisions (of 35), implicit vs declared:

- lattice: 2 vs 3 (Fisher $p > .99$)
- bidirectional lattice: 4 vs 7 ($p = .513$).

The continuous difference $\Delta(A-B)$ (exploratory; not addendum-specified):

- in the lattice, it did not shift (Mann–Whitney $U = 709$, $p = .259$)

- in the bidirectional lattice, it shifted toward Chain A under declared status ($U = 822$, $p = .014$).

The bidirectional shift concerns an input that the baseline model had revised during debate. It remained mostly below the 0.05 alignment threshold, and categorical alignment did not change.

Section S9. Chain-terminal models

Primary-coder chain-terminal regressions, standard condition, with seed fixed effects. Wald p-values for the joint architecture term:

Supplementary Table S18. Chain-terminal regressions with seed fixed effects: Wald p for the joint architecture term (primary coder, standard condition; hallucination models omitted, Methods; post hoc for the primary coder).

Outcome	Architecture Wald p
Evolutionary fidelity	< .001
Completeness	< .001

Pooled models across conditions:

- Architecture: fidelity $p < .001$, completeness $p < .001$.
- Capability condition: fidelity $p .573$, completeness $p .356$.

The primary-coder hallucination regressions, including the node-level mixed model of log hallucination, are omitted because they model primary-coder hallucination magnitudes (Methods, primary-coder parsing). They remain in the OSF code and output release.

Section S10. Seed sentences

Seed 1 (Factual). A 2024 audit of the regional supply chain identified 14 critical vulnerabilities across three distribution hubs, with the highest concentration in cold storage facilities operating below the mandated 34-degree Fahrenheit threshold. The audit estimated that unaddressed vulnerabilities would increase spoilage losses by approximately 23 percent annually, at a projected cost of \$4.2 million to the consortium.

Seed 2 (Normative). The ethics review committee concluded that the proposed clinical trial protocol presents unacceptable risks to vulnerable populations and should not proceed without substantial revision. The committee noted particular concern about the inadequacy of the informed consent procedures and the absence of independent safety monitoring, both of which fall below established standards for research involving older adult participants with mild cognitive decline.

Seed 3 (Narrative). When the lead researcher noticed the anomalous readings at 11:47 PM on a Tuesday, she initially attributed them to sensor calibration drift. It was only after cross-referencing three weeks of archived data that the pattern became clear: the facility had been

leaking trace amounts of a regulated compound into the groundwater since the previous March, undetected by the automated monitoring system.

Seed 4 (Procedural). To initiate the data recovery protocol, the on-call engineer must first isolate the affected partition by disabling write access, then notify the security team before attempting any restoration. Recovery should proceed only from backups created within the previous 72 hours, and all actions must be logged in the incident management system before the engineer completes the incident record.

Seed 5 (Complex: Factual-Narrative). Between January and September of last year, the three largest hospital networks in the region reported a combined 340 cases of a rare post-surgical complication, representing a 280 percent increase over the same period the prior year. Investigators tracing the pattern eventually identified a common thread: all affected patients had received a specific lot of surgical mesh manufactured by a single supplier, which had passed all required quality inspections.

Seed 6 (Complex: Normative-Procedural). The board's policy requires that any AI-generated recommendation affecting more than 500 employees must be reviewed by at least two senior managers before implementation, and that the reviewing managers must document their independent assessments before comparing notes. This requirement exists specifically to prevent anchoring effects in high-stakes decisions, and exceptions may be granted only by the chief executive with written justification.

Seed 7 (Bartlett stimulus, adapted from Bartlett, 1932). One night two young men from Egulac went down to the river to hunt seals. They heard war-cries and a canoe with five warriors appeared. One young man agreed to join the fight though he could not understand why. In the battle he was not hurt, though many died around him. Later he told the people of Egulac what had happened. When the sun rose he fell down, and something black came out of his mouth. He was dead.

Section S11. Leave-one-seed-out stability

Architecture rankings were recomputed with each seed excluded in turn and compared with the full-sample ranking (Spearman ρ , standard condition, primary coder).

Supplementary Table S19. *Leave-one-seed-out stability of primary-coder architecture rankings. Primary factual-retention scores share the production parser; these rank-based results are retained for completeness, but no claim rests on them.*

Outcome	Minimum ρ over the seven exclusions	ρ with Bartlett excluded
Evolutionary fidelity	0.89	0.89
Completeness	0.99	0.99
Factual retention	0.77	0.77
XL-SD	0.94	—

Fidelity and completeness rankings are stable under the exclusion of any single seed. Primary-coder hallucination leave-one-seed-out rankings are not interpreted (Section S16). Outcome magnitudes are not seed-invariant.

Bartlett. With the Bartlett seed excluded, the primary coder's rank-based configuration difference in hallucination remained (Kruskal–Wallis $p < .001$; $< .001$ and $< .001$ under the two parser stress tests). Hallucination magnitudes come from the independent coders:

Supplementary Table S20. Mean terminal hallucination for organizational seeds and the Bartlett seed under the independent coders (standard condition; post hoc). Linear-chain Bartlett medians were 35 under GPT-4.1 and 20 under Gemini.

Architecture	GPT-4.1: organizational	GPT-4.1: Bartlett	Gemini: organizational	Gemini: Bartlett
Linear	6.5	38.0	2.2	16.0
Ring	6.5	38.0	1.3	12.6
Hub	4.0	25.7	0.7	5.6
Hierarchy	4.3	27.7	1.2	6.5
Lattice	5.5	46.0	1.8	18.0
Bidirectional	9.3	49.0	3.7	28.0

Section S12. Cross-lineage divergence under a single delivered output

For hub-and-spoke and the hierarchy, each of the 30 combinations of delivered spoke and delivered lower-tier node is one specification. XL-SD is computed within each specification and then averaged uniformly. The single-output estimand is post hoc for the primary coder and was fixed in the deposited protocol (D4) for the independent coders. The table gives standard-condition results by coder: blocked randomization over seed blocks (Friedman statistic, within-seed permutation, 100,000 draws) and the architecture means.

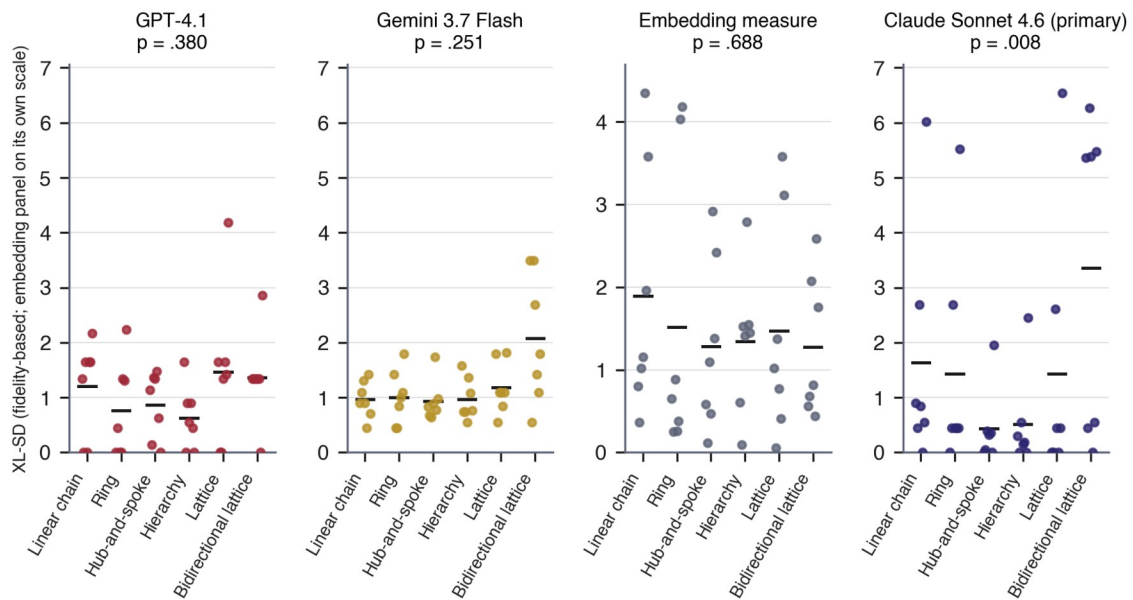
Supplementary Table S21. *Cross-lineage divergence under a single delivered output, by coder and embedding measure.*

Coder	Blocked randomization p	Range over 30 specifications	Asymptotic Friedman p	Kruskal–Wallis p	Linear	Ring	Hub	Hierarchy	Lattice	Bidirectional
Sonnet 4.6	.008	0.00003–0.030	.014	.116	1.63	1.43	0.44	0.52	1.43	3.36
GPT-4.1	.380	.278–.737	.366	.430	1.21	0.76	0.87	0.63	1.46	1.37
Gemini 3.7 Flash	.251	.071–.503	.245	.153	0.96	1.00	0.93	0.97	1.18	2.07
SBERT	.688	—	.662	.933	1.89	1.52	1.28	1.34	1.47	1.27

Under the primary coder:

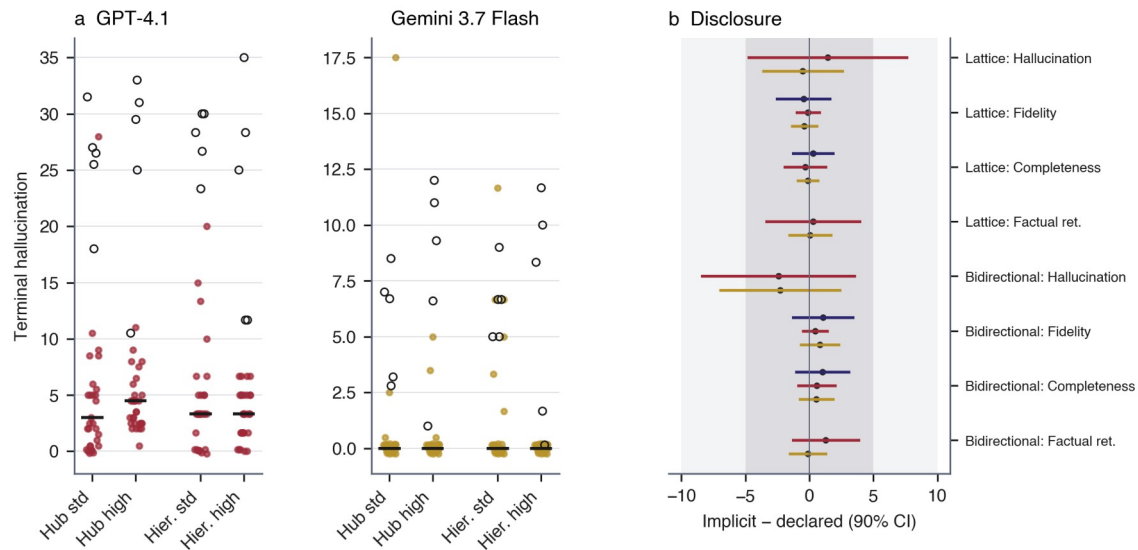
- hub-and-spoke and the hierarchy were the two lowest in 100% of specifications
- the bidirectional lattice was highest in 100%
- the blocked test was significant in every specification, with Bartlett excluded ($p = .007$), and under the asymptotic Friedman test ($p = .014$), but not under the Kruskal–Wallis test that ignores seed blocks ($p = .116$)
- the earlier averaged-terminal definition gave $p = .076$.

Neither independent coder nor the embedding measure reproduced the effect.



Supplementary Fig. S1. The architecture-level XL-SD difference is not reproduced.

Standard-condition XL-SD (single delivered output) by architecture under GPT-4.1, Gemini 3.7 Flash, the embedding measure and the primary coder, each with its blocked randomization p . Points are architecture $\times$ seed cells (seven per architecture) and bars are means. The three LLM-coder panels share a common scale from zero; the embedding panel has its own scale from zero. Only the primary coder detects a difference.



Supplementary Fig. S2. Capability placement and disclosure. (a) Chain-terminal hallucination under the standard and higher-capability conditions for hub-and-spoke and the hierarchy, under the two independent coders (Section S16; Supplementary Table S35). Points are chains (zero values jittered slightly), Bartlett chains are open black rings and bars are medians. Capability placement (H3) is not supported (Section S1). (b) Implicit-minus-declared differences (Section S7) in the lattice and bidirectional lattice with 90% confidence intervals by coder (navy primary, red GPT-4.1, gold Gemini), against equivalence bands at ± 5 and ± 10 points. Primary-coder hallucination and factual-retention intervals are omitted because those magnitudes are not interpreted (Methods).

Section S13. Pooled analyses across conditions (sensitivity)

Capability conditions are only partially crossed with architecture:

- the linear chain and ring were run only in the standard condition
- hub-and-spoke and the hierarchy also have a higher-capability condition
- the lattice and bidirectional lattice also have implicit and declared conditions.

Pooled analyses therefore mix architecture with capability condition and are reported only as sensitivity analyses. The table gives pooled chain-terminal Kruskal–Wallis tests (420 chains) and the rank correlation of pooled with standard-condition architecture means (primary coder).

Supplementary Table S22. Pooled analyses across all conditions (sensitivity; primary coder; hallucination omitted, Section S16).

Outcome	Pooled KW p	Pooled η^2_H	Rank ρ , pooled vs standard
Evolutionary fidelity	< .001	.069	1.00
Completeness	< .001	.112	1.00
Factual retention	.257	.004	0.83

Under the primary coder, pooled XL-SD across 84 cells gave Kruskal–Wallis $p < .001$. The pooled result is stronger than the balanced standard-condition result because it adds capability conditions unevenly across architectures.

Section S14. Seed-blocked concordance

Friedman tests on seed-blocked architecture means (standard condition) and Kendall's W. These tests treat the seven tested messages as blocks and assess whether the configuration difference is consistent across them.

Supplementary Table S23. Seed-blocked Friedman tests and Kendall's W (primary coder, standard condition; post hoc for the primary coder).

Outcome	Friedman p	Kendall W
Hallucination	.057	0.31
Evolutionary fidelity	.023	0.37
Completeness	.007	0.46
XL-SD	.014	0.41

Under the two parser stress tests, the primary coder's hallucination Friedman p is .053 (100 → 35) and .015 (100 → missing).

Supplementary Table S24. Seed-blocked Friedman tests and Kendall's W by coder (standard condition; post hoc for the independent coders).

Outcome	Primary coder p	GPT-4.1 p	GPT-4.1 W	Gemini p	Gemini W
Hallucination	.057	.044	0.33	.281	0.18
Evolutionary fidelity	.023	.001	0.58	.008	0.44
Completeness	.007	.010	0.43	.145	0.23

Fidelity is detected under all three coders, completeness under two, and hallucination is the least consistent across messages.

Section S15. Hub-and-spoke terminal outputs

Hub-and-spoke has ten terminal outputs per chain, each two transformations from the seed. The common terminal rule uses their mean for the fidelity, completeness, factual-retention and hallucination outcomes. Divergence uses a single delivered spoke, averaged over specifications (S12). The stored spoke positions (2–11) are a storage convention and are not used as a depth axis.

Section S16. Independent recoding

Protocol. The recoding protocol, frozen sample, instruments, model versions, settings, analyses and interpretation rules were deposited on OSF before execution (not a preregistration). Coders:

- OpenAI gpt-4.1-2025-04-14: temperature 0.1, 16 output tokens for Instrument A, JSON output for Instrument B

- Google gemini-3.7-flash: temperature 1.0 (provider default), thinking level "low", room for thinking tokens, JSON output for Instrument B.

Both coders ran on standard API tiers. No Anthropic model was used as an independent coder. Each rating was a single call. The analyses use the first replicate, and replicates 2–3 estimate repeatability.

Sample.

- 2,581 unique outputs (44% of all recorded outputs): all 2,520 outputs required by the terminal and depth-referenced analyses specified in the frozen protocol, plus 61 additional outputs contributed uniquely by the earlier 150-pair validation sample
- 517 rated three times
- 2,207 eligible for Instrument B (organizational seeds).

Condition labels were hidden from the coders.

Execution.

Supplementary Table S25. *Execution of the independent recoding.*

Coder	Coding jobs	Successful	Missing
GPT-4.1	17,545	17,545	0
Gemini 3.7 Flash	17,545	17,545	0

Every returned model identifier matched the pinned version. The execution record also covers retries, spend and quota interruptions (published with the outputs).

Deviations from the protocol, all recorded in the deposited deviation log:

1. The frozen analysis scripts omitted four analyses that the protocol text specified (agreement by tier, all replicate pairs, the full depth-contrast family, and origin tracing for anchored shifts). A supplementary script implementing the protocol text was written and validated against the primary coder's own scores before any substantive recode results were inspected.
2. The Gemini run was twice interrupted by the provider's daily request quota and resumed, without change, after reset.
3. Under-specified interpretation rules ("comparable magnitude", ordering and consistency diagnostics) were operationalized before any substantive recode results were inspected. The ordering and consistency thresholds are diagnostic classifications, not verdicts.
4. After the results were produced, the protocol-prescribed interpretive wording for factual retention (rule 9) was withdrawn because its comparative clause was not supported across coders. No statistic, threshold, or analysis changed.

Row-level agreement and repeatability (D1–D2).

Supplementary Table S26. *Row-level agreement with the primary coder and within-coder repeatability of the independent coders.*

Outcome	ρ with Sonnet: GPT-4.1	ρ with Sonnet: Gemini	Replicate ρ : GPT-4.1	Replicate ρ : Gemini	Exact agreement: GPT-4.1	Exact agreement: Gemini
Hallucination	.74	.69	.96–.97	.95–.97	85%–89%	88%–89%
Evolutionary fidelity	.57	.57	.82–.84	.79–.82	86%–88%	54%–60%
Completeness	.62	.36	.88–.90	.74–.78	84%–86%	94%–95%
Factual retention	.20	–.17	.88–.93	.81–.85	88%–92%	93%–94%

Primary-coder parser diagnostics and stress tests (post hoc). The production pipeline converted each primary-coder reply to a number by deleting every non-digit character and capping the result at 100. Raw replies were not retained. A reply such as "35/100" would be stored as 100, and other replies containing several digit sequences could be misread at other values, while plain integers parse correctly. The independent coders' replies were parsed strictly, and GPT-4.1's strict and production parses agree on every row.

Diagnostics for primary-coder hallucination: 398 of 5,880 scores are exactly 100 and 0 lie between 36 and 94 (the largest score below 95 is 35). Of the 100-scored outputs, 78% come from the Bartlett seed and 30% have primary fidelity of 96 or more. Of the 228 100-scored outputs in the recode sample, GPT-4.1 gave a median of 35 (maximum 70) and Gemini a median of 10 (maximum 40).

Primary-coder hallucination magnitudes are therefore not interpreted anywhere in the paper. Two declared stress tests rerun the affected primary-coder rank-based tests with every score of 100 set to 35, or treated as missing. They bound the artefact and are not corrected estimates.

Supplementary Table S27. *Primary-coder hallucination: rank-based tests under the two parser stress tests (standard condition; post hoc).*

Test	As recorded	100 → 35	100 → missing
Kruskal–Wallis p	.005	.003	.004
Seed-stratified omnibus p	< .001	< .001	< .001
Kruskal–Wallis p, Bartlett excluded	< .001	< .001	< .001
Holm-surviving pairs (of 15)	2	2	2
of which involve the bidirectional lattice	2	2	2
Seed-blocked Friedman p	.057	.053	.015
Capability placement, hub-and-spoke (seed-stratified p)	.123	.785	.020
Capability placement, hierarchy (seed-stratified p)	.045	.043	.840
Matched-depth contrasts surviving Holm, critical	1	1	0
Matched-depth contrasts surviving Holm, message-carrying route	1	1	0
Hierarchy mid-tier synthesis at depth 2, critical (Holm p)	.019	.019	.612

The configuration omnibus, the Bartlett-excluded test and the two corrected pairs (hub-and-spoke and the hierarchy, each against the bidirectional lattice) are unchanged in substance. The capability-placement and mid-tier contrasts are not stable and are not interpreted.

The same parser applied to the other primary outcomes. Only 1 primary fidelity score and 4 completeness scores are exactly 100, so neither measure shows the ceiling-spike signature seen

for hallucination. Because the raw replies were not retained, however, other parser-induced misreadings cannot be ruled out. Factual retention has 1,688 ceiling scores, so genuine full-retention ratings cannot be distinguished from parser-induced ceiling values. Primary-coder factual-retention magnitudes therefore carry no claim. No correction is made.

Post hoc descriptive additions (not pre-specified). Two sets of values were computed from the frozen recoding outputs, using the same rows and definitions as the table above. There are no inferential tests. The first is row-level agreement between the two independent coders. The second is the association between chain-terminal factual retention and hallucination for each coder.

Supplementary Table S28. *Row-level agreement between the two independent coders (Spearman ρ , primary replicate).*

Outcome	ρ GPT-4.1 vs Gemini
Hallucination	.71
Evolutionary fidelity	.52
Completeness	.30
Factual retention	.53

Supplementary Table S29. *Pearson r between chain-terminal factual retention and hallucination, by independent coder. The primary coder's association is not reported because both of its scores share the production parser (Section S16).*

Coder	Standard condition	Standard, Bartlett excluded	All 12 conditions
GPT-4.1	-.86	-.38	-.85
Gemini 3.7 Flash	-.67	-.04	-.59

The association is sensitive to the Bartlett seed, so the main text does not use it as evidence about marker retention.

Configuration comparison by coder (D3). Standard-condition chain-terminal means, Kruskal–Wallis and seed-stratified omnibus p, number of Holm-surviving pairs (those involving the bidirectional lattice in parentheses).

Supplementary Table S30. *Standard-condition configuration comparison by coder. Primary-coder hallucination and factual-retention means are not shown (—) because those magnitudes are not interpreted; its rank-based hallucination tests survive both parser stress tests (see above). Primary factual-retention scores share the production parser; these rank-based results are retained for completeness, but no claim rests on them.*

Outcome	Coder	Linear	Ring	Hub	Hierarchy	Lattice	Bidirectional	KW p	Seed-stratified p	Corrected pairs (bidirectional)
Hallucination	Sonnet 4.6 (primary)	—	—	—	—	—	—	.005	< .001	2 (2)
Hallucination	GPT-4.1	11.00	11.00	7.06	7.67	11.29	15.00	.140	< .001	3 (3)
Hallucination	Gemini 3.7 Flash	4.14	2.94	1.39	1.92	4.14	7.14	.449	.014	0 (0)
Evolutionary fidelity	Sonnet 4.6 (primary)	94.83	95.86	96.61	96.54	95.26	91.49	.036	< .001	4 (4)
Evolutionary fidelity	GPT-4.1	95.63	96.09	96.72	96.51	95.51	94.60	< .001	< .001	7 (3)
Evolutionary fidelity	Gemini 3.7 Flash	97.26	97.31	97.93	97.74	97.03	95.06	.002	< .001	3 (3)
Completeness	Sonnet 4.6 (primary)	96.91	97.40	97.94	98.09	96.40	93.80	< .001	< .001	9 (5)
Completeness	GPT-4.1	96.34	96.11	97.20	97.00	95.86	94.26	.005	< .001	5 (5)
Completeness	Gemini 3.7 Flash	99.57	99.86	99.93	99.88	99.46	98.31	.040	.004	1 (1)
Factual retention	Sonnet 4.6 (primary)	—	—	—	—	—	—	.640	.019	3 (3)
Factual retention	GPT-4.1	92.57	93.34	94.36	94.10	92.86	92.14	.987	.404	0 (0)
Factual retention	Gemini 3.7 Flash	98.66	99.51	99.66	99.51	98.83	98.09	.428	.046	0 (0)

Pairwise tests (D3b).

Supplementary Table S31. Hallucination: Holm-adjusted seed-stratified rank permutation p over the 15 architecture pairs, by coder. For the primary coder only the Holm significance status ($p < .05$) is shown; the same two pairs are significant under both parser stress tests (Section S16).

Pair	Sonnet 4.6 (significant after Holm)	GPT-4.1	Gemini 3.7 Flash
Linear vs Ring	no	> .99	> .99
Linear vs Hub	no	> .99	.361
Linear vs Hierarchy	no	> .99	> .99
Linear vs Lattice	no	> .99	> .99
Linear vs Bidirectional	no	.092	> .99
Ring vs Hub	no	.542	.113
Ring vs Hierarchy	no	.328	> .99
Ring vs Lattice	no	> .99	> .99
Ring vs Bidirectional	no	.109	.733
Hub vs Hierarchy	no	> .99	> .99
Hub vs Lattice	no	> .99	.361
Hub vs Bidirectional	yes	< .001	.068
Hierarchy vs Lattice	no	> .99	> .99
Hierarchy vs Bidirectional	yes	< .001	.367
Lattice vs Bidirectional	no	.031	> .99

Supplementary Table S32. Evolutionary fidelity: Holm-adjusted seed-stratified rank permutation p over the 15 architecture pairs, by coder.

Pair	Sonnet 4.6	GPT-4.1	Gemini 3.7 Flash
Linear vs Ring	.696	.932	> .99
Linear vs Hub	.098	.033	.085
Linear vs Hierarchy	.098	.032	.172
Linear vs Lattice	.696	> .99	> .99
Linear vs Bidirectional	.242	.075	.034
Ring vs Hub	> .99	.117	.155
Ring vs Hierarchy	> .99	.117	.140
Ring vs Lattice	> .99	.939	> .99
Ring vs Bidirectional	.006	.003	.098
Hub vs Hierarchy	> .99	> .99	> .99
Hub vs Lattice	> .99	.018	.061
Hub vs Bidirectional	< .001	< .001	< .001
Hierarchy vs Lattice	> .99	.033	.155
Hierarchy vs Bidirectional	.001	< .001	.001
Lattice vs Bidirectional	.004	.075	.100

Supplementary Table S33. Completeness: Holm-adjusted seed-stratified rank permutation p over the 15 architecture pairs, by coder.

Pair	Sonnet 4.6	GPT-4.1	Gemini 3.7 Flash
Linear vs Ring	> .99	> .99	> .99
Linear vs Hub	.026	> .99	> .99
Linear vs Hierarchy	.014	> .99	> .99
Linear vs Lattice	> .99	> .99	> .99
Linear vs Bidirectional	.005	.008	.090
Ring vs Hub	.664	> .99	.776
Ring vs Hierarchy	.032	> .99	> .99
Ring vs Lattice	> .99	> .99	> .99
Ring vs Bidirectional	.003	.011	.016

Pair	Sonnet 4.6	GPT-4.1	Gemini 3.7 Flash
Hub vs Hierarchy	> .99	> .99	> .99
Hub vs Lattice	.124	> .99	> .99
Hub vs Bidirectional	< .001	< .001	.323
Hierarchy vs Lattice	.026	.838	> .99
Hierarchy vs Bidirectional	< .001	< .001	.091
Lattice vs Bidirectional	.018	.023	.379

Supplementary Table S34. *Factual retention: Holm-adjusted seed-stratified rank permutation p over the 15 architecture pairs, by coder. Primary factual-retention scores share the production parser; these rank-based results are retained for completeness, but no claim rests on them.*

Pair	Sonnet 4.6	GPT-4.1	Gemini 3.7 Flash
Linear vs Ring	> .99	> .99	.766
Linear vs Hub	> .99	> .99	.976
Linear vs Hierarchy	> .99	> .99	.976
Linear vs Lattice	> .99	> .99	> .99
Linear vs Bidirectional	.112	> .99	> .99
Ring vs Hub	> .99	> .99	> .99
Ring vs Hierarchy	> .99	> .99	> .99
Ring vs Lattice	> .99	> .99	> .99
Ring vs Bidirectional	.433	> .99	.108
Hub vs Hierarchy	> .99	> .99	> .99
Hub vs Lattice	> .99	> .99	> .99
Hub vs Bidirectional	.009	> .99	.976
Hierarchy vs Lattice	> .99	> .99	> .99
Hierarchy vs Bidirectional	.034	.103	.976
Lattice vs Bidirectional	.042	> .99	.976

Divergence (D4). See S12.

Depth-referenced contrasts (D5) and debate divergence (D6). See S6.

Capability placement sensitivity (D7). Chain-terminal hallucination, standard vs higher-capability condition, seed-stratified permutation p .

Supplementary Table S35. *Capability placement sensitivity by independent coder (chain-terminal hallucination, standard versus higher-capability condition). The primary coder's placement contrasts are not stable under the parser stress tests (see above).*

Architecture	Seeds	GPT-4.1: standard vs high (p)	Gemini 3.7 Flash: standard vs high (p)
Hub	all seeds	7.06 vs 7.40 (.760)	1.39 vs 1.40 (> .99)
Hub	Bartlett excluded	3.95 vs 4.33 (.756)	0.68 vs 0.30 (.841)
Hierarchy	all seeds	7.67 vs 5.95 (.068)	1.92 vs 0.90 (.049)
Hierarchy	Bartlett excluded	4.33 vs 3.22 (.143)	1.17 vs 0.00 (.011)

Section S17. Proposition-anchored epistemic status (Instrument B)

Codebook. Two anchors per organizational seed. Each is a proposition from the seed with its epistemic status in the source. For each anchor present in an output, the coder judged whether its status was preserved, strengthened, weakened, altered or uncertain. A status coded strengthened, weakened or altered counts as a shift; uncertain does not. The full prompt and codebook are in the OSF deposit.

Supplementary Table S36. *Instrument B codebook: source propositions and their epistemic status in the seed.*

Anchor	Seed	Source proposition	Source epistemic status
S1a	1	Unaddressed vulnerabilities would increase spoilage losses by approximately 23 percent annually.	conditional projection (estimated by the audit; contingent on vulnerabilities remaining unaddressed)
S1b	1	The \$4.2 million is a projected cost to the consortium.	projected, conditional cost
S2a	2	The ethics review committee concluded that the protocol presents unacceptable risks to vulnerable populations.	attributed judgement of a named body
S2b	2	The trial should not proceed without substantial revision.	conditional recommendation (proceeding is allowed after substantial revision)
S3a	3	The researcher initially attributed the anomalous readings to sensor calibration drift.	initial hypothesis later superseded
S3b	3	The facility had been leaking a regulated compound into the groundwater since the previous March, undetected.	established finding after cross-referencing archived data
S4a	4	Recovery should proceed only from backups created within the previous 72 hours.	restricted procedural requirement
S4b	4	The engineer must isolate the partition and notify the security team before attempting restoration.	mandatory, ordered prerequisites
S5a	5	All affected patients had received a specific lot of surgical mesh from a single supplier (a common thread).	association / commonality identified by investigators
S5b	5	The mesh lot had passed all required quality inspections.	stated fact (concessive)
S6a	6	AI-generated recommendations affecting more than 500 employees require review by at least two senior managers.	conditional threshold requirement
S6b	6	Exceptions may be granted only by the chief executive with written justification.	limited permission

Agreement between the independent coders. The pre-specified credibility rule for any prevalence statement is between-coder shift agreement of Gwet's $AC1 \geq 0.60$. Where shifts are rare, $AC1$ can be high while κ is near zero. Agreement then means agreement that shifts are rare.

Supplementary Table S37. *Instrument B: shift rates and between-coder agreement by anchor.*

Anchor	Rows with anchor present (both coders)	Shift rate GPT-4.1	Shift rate Gemini	Cohen κ	Gwet $AC1$	Credibility rule ($AC1 \geq 0.60$)
S1a	371	25.3%	42.6%	.43	.52	not met
S1b	371	21.0%	13.2%	.30	.72	met
S2a	363	6.1%	5.0%	.79	.98	met
S2b	362	11.6%	79.3%	-.01	-.45	not met
S3a	372	0.3%	0.0%	.00	1.00	met
S3b	372	18.5%	0.3%	.02	.78	met
S4a	365	19.7%	36.7%	.54	.68	met
S4b	365	9.9%	3.6%	.38	.91	met
S5a	372	53.5%	19.9%	.22	.24	not met
S5b	372	0.3%	0.0%	.00	1.00	met
S6a	364	0.5%	2.2%	-.01	.97	met
S6b	364	0.3%	0.3%	1.00	1.00	met

Agreed terminal shifts (descriptive only). For anchors meeting the rule, the table gives the share of terminal outputs in which both coders marked a shift, by architecture. The shares pool all capability conditions, and the denominators differ sharply. Terminal outputs per anchor are 5 for the linear chain, 5 for the ring, 100 for hub-and-spoke, 30 for the hierarchy, 15 for the lattice and 15 for the bidirectional lattice. The table therefore does not estimate architecture differences or architecture-specific prevalence, and zeros in the linear chain and ring are not informative.

Supplementary Table S38. *Instrument B: share of terminal outputs with an agreed shift, by anchor and architecture (all conditions pooled; descriptive).*

Anchor	Linear	Ring	Hub	Hierarchy	Lattice	Bidirectional
S1b	0.00	0.00	0.06	0.03	0.13	0.20
S2a	0.00	0.00	0.12	0.03	0.07	0.13
S3a	0.00	0.00	0.00	0.00	0.00	0.00
S3b	0.00	0.00	0.00	0.00	0.00	0.07
S4a	0.00	0.00	0.27	0.20	0.13	0.20
S4b	0.00	0.00	0.00	0.00	0.00	0.13
S5b	0.00	0.00	0.00	0.00	0.00	0.00
S6a	0.00	0.00	0.00	0.00	0.00	0.00
S6b	0.00	0.00	0.00	0.00	0.00	0.00

Origin. For each chain with an agreed terminal shift, the earliest recorded upstream output at which both coders marked a shift was identified. In the documented Seed 1 case (main-text Fig. 3), the agreed shift is on anchor S1b, and it originates at the post-debate state of Chain B. Anchor S1a, the projected increase in spoilage losses, did not meet the credibility rule. The full origin table is in the OSF deposit.