Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When Does Hallucination Propagate? Reconciling Six Formalizations of Error Spread in Multi-Agent LLM Systems

Zhang Shaoxuan

2023113033@stu.hit.edu.cn

Harbin Institute of Technology https://orcid.org/0009-0001-4103-6099

Systematic Review

Keywords: hallucination propagation, multi-agent systems, verification, aggregation loss

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11536507/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

When Does Hallucination Propagate? Reconciling Six Formalizations of Error Spread in Multi-Agent LLM Systems

Shaoxuan Zhang
Faculty of Computing, Harbin Institute of Technology, Harbin, China
Corresponding author — 2023113033@stu.hit.edu.cn · ORCID 0009-0001-4103-6099

Abstract

Multi-agent LLM systems are deployed on the assumption that agents checking one another reduce hallucination. Two studies of the same linear topology contradict it in opposite directions, and both are right: one measures prevalence, the other detectability, and a claim can become rarer and less visible at once.

Whether a false claim grows is set by two structural variables: how correlated the agents' errors are, and whether any channel supplies evidence independent of them. Topology bounds the number of carriers a chain can produce; it says nothing about the other channel.

Six formalisms proposed in the past year describe this process. Each defines a state space, a transition rule and a measurable quantity; none is stated in terms another can absorb. Taking the travelling claim as the unit of analysis, we show that the six are projections of one process onto different observables.

In a purposive sample of 42 papers, 10 of 32 included studies report a propagation-level metric, 9 report a verification channel independent of the agents, and 7 report a graph with a recoverable out-degree. The two features never co-occur: all nine studies with such a channel report attenuation, and both studies that report amplification lack one. One prediction runs against design intuition: raising the initial diversity of agents' answers increases false adoption. We specify the experiment that separates a causal reading from an artefact, and a reporting standard — carrier set, aggregation loss, out-degree distribution, comparator — under which the six become comparable.

Keywords: hallucination propagation; multi-agent systems; verification; aggregation loss

1. Introduction

Multi-agent systems built on large language models [1, 2, 3, 4] are now used where a factual error is expensive: clinical triage, contract review, financial analysis, screening of scientific literature. The arrangement rests on an intuitive claim. An agent that reads another agent's output will catch some of its mistakes, so a group should be at least as reliable as its most reliable member. The assumption is the one behind debate protocols [5, 6] and consistency voting [7]: independent samples are treated as independent errors. Recent measurements do not support it. Collaboration buys task performance more dependably than it buys factual reliability [8], and under some configurations it makes hallucination worse [9]. Hallucination itself is by now well catalogued [10, 11]; what is not catalogued is what a hallucinated claim does once it starts to move.

Two 2026 studies make the difficulty concrete. Both follow a hallucinated claim through a linear pipeline in which each agent feeds exactly one successor. In a three-agent cascade spanning ten knowledge domains, the normalized hallucination score falls from 0.422 at the first agent to 0.272 at the third, an amplification factor of 0.644: net attenuation [12]. In a four-agent financial pipeline with 346 injected hallucinations, detection of those hallucinations falls from 72.0% at the first stage to 50.9% at the fourth, and 23.7% are never detected at all: net amplification [13]. Both pipelines are chains. Neither result is wrong.

The two studies measure different things. One asks how much of the generated text lacks support; the other asks whether a reader can still tell that it does. A claim can become rarer and less detectable at the same time, and a literature that reports only one of these quantities will appear to contradict itself. This is not a defect of either paper. It is what happens when a phenomenon is measured along axes that are never brought into the same frame.

That frame is what the field currently lacks, and the absence shows in how fast its formalisms multiply. Six have appeared within roughly a year. Hallucination has been modelled as a first-order Markov process over four states — raw fact, derived computation, narrative prose, editorially approved conclusion — with measured per-boundary escape probabilities of 24.6%, 48.3% and 89.3% [13]. It has been modelled as a time-evolving network process on directed communication graphs, yielding an effective reproduction number that a defence lowers from 1.08 to 0.81 [14]. It has been modelled as a claim-level stochastic cascade with an amplification factor, a relative propagation rate and a cumulative change term [12]; as a divergence of internal knowledge states between agent pairs, scored by a context divergence index [15]; as confidence-normalized pooling over a communication topology, with a conformity index and a time-to-consensus [16]; and as a cross-channel causal influence tensor monitored through its spectral dynamics [17]. Each contributes a state space, a transition rule and at least one measurable quantity. None is stated in terms another can absorb.

The disagreements extend past vocabulary. One position paper argues that safety properties in agentic systems are set by interaction topology rather than by model scale or alignment, and names information cascades among three topology-driven pathologies [18]. A controlled clinical study reaches the opposite emphasis: across four systems, four models and three tasks, the model rather than the topology or the agent count determines both the level of hallucination and the direction of any change [8]. Both are empirical. They cannot both be right in the form stated.

Existing surveys do not resolve this, because they are organised along a different axis (Table 1). One provides the first taxonomy of hallucination in LLM-based agents, structured by the stages of a single agent's workflow [19]. Another organises multi-agent reliability as a four-stage progression — lay the capability foundation, integrate agents, find faults through attribution, evolve — and treats error propagation as the risk that tighter coordination introduces [20]. A third arranges mitigation by application setting: retrieval, reasoning, agentic deployment [21]. A fourth catalogues collaboration mechanisms — who talks to whom, in what pattern — without following the error that travels through them [22]. Each is useful for what it covers. None takes the travelling claim as its unit, and two of them already use the stage-table device, which makes a further five-stage taxonomy a poor way to distinguish a new survey.

What is missing is narrower than a framework and more useful than a list. It is a statement of when the sign of propagation is positive, together with a measurement standard that lets six incompatible formalisms be compared. We address both. Our organising principle is the claim rather than the agent or the stage: a hallucination is a proposition that acquires or loses support as it travels, and every mechanism in the literature can be read as an operation on that support.

We make four contributions. First, we give a conditional account. Whether a false claim grows is, on our account, set by two structural variables — the correlation pressure among agents, and the independence of whatever verification channel exists — while the topology sets whether amplification can compound rather than whether it occurs. We propose that the topology-versus-model dispute is better read as a question about which variable dominates on which side of a threshold. We state that reading as a conjecture, because no published experiment separates the two variables while holding the other fixed (§3.8). Second, we separate four axes that the literature routinely conflates: what is harmed (false claims gaining support, or true claims being discarded), in what form (breadth or strength), at which stage, and under which structural condition (§3.2–3.6). Third, we code 42 papers against these axes — a purposive sample, not an exhaustive screen — and report what the map shows: 10 of 32 included studies report any propagation-level metric, 9 report an independent verification channel, and 7 report a communication graph from which an out-degree can be recovered, though 15 state an agent count. The two features never co-occur: all nine studies with an independent channel report attenuation, and both studies that report amplification lack one. The two quantities the framework needs most — an effective reproduction number and an aggregation loss — each appear in exactly one paper (§2). Fourth, we map interventions onto parameters, prove that no aggregation rule removes the aggregation loss uniformly over candidate pools (Proposition 3), and conclude that only an independent channel lowers both outcome channels at once — which is why recovery, and not prevention alone, has to be a first-class component of the design (§5).

The remainder of the paper follows that order. §2 states the coding protocol and what the corpus shows. §3 builds the formal vocabulary, defines the two propositions, and places the six existing formalisms inside it. §4 covers the stages in turn, §5 the interventions, and §6 sets out the measurement and recovery agenda that the gaps imply.

Table 1. What the adjacent surveys organise around

Each of the three surveys in this table answers a question this paper does not, and does so at a depth this paper does not attempt. The table marks where the questions differ, not where the answers are weaker.

	Lin et al. [19]	Qi et al. [20] (LIFE)	Li et al. [21]	This survey
Unit of analysis	the agent	the system	the application	the claim
Organising axis	stages of one agent's workflow	four-stage capability progression: lay, integrate, find faults, evolve	deployment setting: RAG, reasoning, agentic	the stage a claim has reached
Question it answers	what kinds of hallucination an agent can have, and what causes them	how a multi-agent system matures, and where the stage boundaries fail	which mitigation applies in which setting	when the sign of propagation is positive, and how to compare the accounts that disagree
Coverage of inter-agent propagation	named as a consequence; not the axis	treated as the risk that tighter coordination introduces	not the axis	the axis
Distinguishes false-claim gain from true-claim loss	no	no	no	yes (§3.5)
Distinguishes a stage from a regime	n/a	stages	n/a	cascading as regime, §4.5
Treats recovery as a component	as detection and mitigation	as system-level self-evolution	as mitigation	six actions with separate metrics, §5.3
Proposes a reporting standard	no	no	no	§6.5
Reconciles competing formalisms	not applicable	not attempted	not applicable	six formalisms, §3.9

Three differences carry the paper's claim to a distinct position.

The first is the unit. A survey organised by agent, by system or by application can describe what each component does; it cannot track a proposition through a system, because the proposition is not an object in its taxonomy. The two chains in §1 are invisible from an agent-level or system-level view: in both cases the agents are ordinary, the architecture is a pipeline, and the only thing that distinguishes the studies is what happened to a claim as it moved.

The second is the outcome split. None of the three separates a false claim gaining support from a true claim being discarded, and §4.5 argues that a framework which does not separate them will report the second as an improvement, because the discarded claims never appear in a hallucination count.

The third is the treatment of disagreement. The three surveys synthesise literatures that broadly agree within their own scope. This one synthesises a literature that does not, and its contribution is a vocabulary in which the disagreement resolves into different measurements of the same process rather than a contest between findings.

2. Corpus, coding, and what the coded set shows

This section states where the evidence in the rest of the paper comes from. It is short on purpose: a survey that reconciles other people's numbers owes an account of its own. We report the retrieval, the selection rule, and the four coding decisions that changed the data most. We then give what the coded set shows, and we state plainly what it cannot support.

We do not report a PRISMA flow, because our sample was not drawn by exhaustive screening of all retrieved records. It was drawn purposively, by quota, from the categories the framework needs. The proportions below therefore describe the coded set, not the field. §2.6 says what that costs.

2.1 Retrieval

Two searches were run against the arXiv API on 2026-10-27, restricted to records from 2022-11 onward — the window in which multi-agent LLM work exists — and deduplicated by base identifier.

The first track targeted explicit hallucination and factual-error language combined with multi-agent vocabulary. It returned 432 unique records across ten Boolean queries. The second track targeted the literatures that discuss propagation under other names: failure and reliability, failure attribution, agent memory, conformity and sycophancy, and propagation theory. It returned 908 records, of which 814 were new.

The second track exists because of a recall failure we could observe directly. *Why Do Multi-Agent LLM Systems Fail?* [23] — a 1,600-trace annotated taxonomy with inter-annotator agreement of $\kappa = 0.88$ — was not returned by any first-track query. Its abstract contains no hallucination, factual or faithfulness vocabulary, so a search defined by the phenomenon's name cannot find the work that studies the phenomenon under a different name. Any single-track search of this literature will have the same blind spot.

2.2 Sampling

From 1,246 unique records we coded 42 rows: 30 drawn by quota and 12 from the formalization search.

The quota was set by which stage of the framework needed evidence: amplification 8, consolidation 5, transmission 5, attribution 5, recovery 4, and 3 boundary cases whose inclusion was doubtful by design. Candidates within each quota were ranked by how many independent queries retrieved them, on the assumption that a record found by several differently-worded searches is more central than one found by a single phrasing. The remaining 12 are the works identified as formalizations (§3.1) plus the excluded near-misses discussed in §3.9.

This is a purposive sample. Its virtue is that it covers the mechanisms the argument needs. Its defect is that it cannot estimate a field-level rate, and we do not use it to. When §2.5 says that 10 of 32 included studies report a propagation-level metric, the denominator is the coded set.

2.3 Inclusion, and what the rule excluded

A row was included if the work (i) involves two or more LLM instances exchanging information, (ii) concerns factuality, error, or the detection and repair of either, and (iii) describes the system well enough to code its structure.

The third condition required a ruling on security work, which we had excluded from scope. We include a security paper when it *formalizes spread* and exclude it when it only reports attack success. A cascade-monitoring framework that estimates directed influence across channels qualifies [17]; a memory-poisoning attack that reports only attack success rate does not [24]. This keeps the excluded object — security as a topic — separate from the included one, which is a model of how influence moves.

Ten of the 42 rows failed inclusion. Eight describe a single agent: seven are memory systems [25, 26, 27, 28, 29, 24, 30] and one evaluates a single agent's trajectory [31]. One is not an LLM system at all — a formation-control paper using "state hallucination" in its control-theoretic sense [32]. One decomposes a single agent's tool-use policy into modules and studies propagation between modules [33].

The concentration in the memory group is a finding rather than an inconvenience. Seven of the ten exclusions are memory work, and every one of them is single-agent. Consolidation, as a stage, currently has almost no multi-agent evidence: the mechanisms that make a stored error durable have been studied thoroughly, and studied on systems with one agent. §4.4 treats this as a gap to be bridged rather than a literature to be summarised.

2.4 Four coding decisions that changed the data

Each of these moved rows between categories, so each is stated rather than buried.

Verification is tested by provenance, not by label. A procedure counts as a verification channel only if its inputs are disjoint from the claim's ancestry (§3.3). Under this test several self-described verification procedures code as absent, including one that exchanges compressed state summaries between agents before joint reasoning [15]. A different model reviewing an output also fails the test: heterogeneity changes how the judgement is computed, not what it is conditioned on.

Direction is coded only against a stated comparator. "Amplifies" and "attenuates" are meaningless without one. We record the comparator in a separate field. In the five-paper pilot the comparison was explicit in the three papers that made a directional claim, and simply not applicable in the two that made none, so the field was cleaner than expected. We keep it separate regardless, because attenuation relative to the first hop and attenuation relative to a single-agent baseline are different findings that the word "attenuates" conceals.

"Topology" is treated as an overloaded term. It refers in some work to the agent communication graph and in others to a knowledge graph over domain concepts. We code only the former. One evaluation framework was excluded on this ground alone [31]: its topology-aware analysis is over scientific concept graphs, and folding it into the agent-topology axis would have inflated that axis with work that does not belong to it.

Unknown is recorded as unknown. Where a field could not be determined from the available text we recorded `UNK` rather than inferring a plausible value. The consequence is visible in §2.5: out-degree is unknown for 25 of 32 included studies. That number is not missing data in the ordinary sense; §2.5 reports the audit that tested it.

2.5 What the coded set shows

Table 2. What the coded set reports. Counts over the 32 included studies.

Quantity	Count
Included studies	32
Report any propagation-level metric	10 (31%)
Report an independent verification channel	9 (28%)
State an agent count	15 (47%)
Report a graph from which out-degree is recoverable	7 (22%)
Report an effective reproduction number	1
Report an aggregation loss	1

The reporting rates are lower than the field's output volume would suggest, and the gap between the third and fourth rows is itself a result. Roughly half the corpus states how many agents it used; under a quarter states a communication graph from which an out-degree can be recovered. Counting agents is a natural thing to report. Specifying out-degree is not, even though the branching argument in §3.8 makes it the quantity that decides whether amplification can compound at all.

We checked that this gap is not an artefact of our coding depth. Twenty-six included studies had no recoverable out-degree from their abstracts. We retrieved the full text of each and searched the method sections for agent counts and communication structure. The audit recovered an agent count for five of the twenty-six and an out-degree for none. Two of the five are worth naming because they change rows: the clinical study reports hallucination rates at *three and seven* agents across three topologies [8], and the collective-hallucination study uses a heterogeneous *six*-agent system across directed graphs [14]. Neither states an out-degree, and the second varies it by topology by construction. The 22% therefore describes the literature. A full-text pass would not raise it.

The last two rows are the ones that matter for the framework, because a claim about a threshold requires a quantity that can cross it, and in this corpus exactly one paper reports such a quantity. Appendix A reports the bounds these counts admit, and the reliability study that would replace them.

Crossing verification against direction gives a cleaner result than we expected. Among the 24 included studies that make a directional claim:

Table 3. Verification channel against reported direction. The 24 included studies that make a directional claim.

	attenuate	mixed	amplify
Independent verification present	9	0	0
Independent verification absent	5	6	2
Unknown	1	1	0

The split is complete and it is an absence of co-occurrence rather than a ratio: every study with an independent channel reports attenuation, and neither study reporting amplification has one. We did not choose the studies to produce this. Three readings are available and we cannot yet separate them. The intended one is causal, per Proposition 1. A second is that mitigation papers add retrieval because retrieval is the intervention they are testing, and mitigation papers are the ones that report improvement. A third is that a study which measures propagation has no reason to build a verifier, so the two practices simply do not co-occur. The second and third readings are about research fashion and division of labour rather than about mechanism, and only an experiment that crosses the two factors would distinguish them. That experiment is the first item in §6.

Two further patterns in the coded set bear on §3.

Seven included studies use a chain topology and none reports amplification in carrier count: six report attenuation and one reports both directions [13], because it measures detectability as well as prevalence. That is what Proposition 2b predicts. The chain constrains how many carriers a claim can produce; it says nothing about how hard the claim becomes to see.

Direction is coded as each paper's headline, not as a channel-specific measurement, because no paper in the corpus reports both channels. This is the single largest limitation of the coded set and the reason §3.5 insists on the FP/FN split: the field's own labels cannot express the distinction that the data appear to require.

2.6 Threats to validity

Single coder. All 42 rows were coded by one person. No inter-rater agreement is reported, and the κ values that a two-coder design would produce are unknown. The three fields that carry the argument — verification, topology, direction — are the ones most exposed. Appendix A bounds the damage: it replaces the verification proportion with the range the unresolved cells admit (28 to 38%), shows that the central split is invariant to verification-coding error in the attenuating group and breaks only if an amplifying study turns out to have a channel, and reports the result of the reliability study that was run against a second coder.

Abstract-level. Three rows were read in full — the two formalizations whose criteria needed checking at method level and one cascade study. The remaining 39 rest on abstracts; for eight of those the abstract was explicit enough to code at high confidence, and the rest are marked

medium or low. Fields such as persistence are frequently `UNK` because the abstract does not say, which is a different problem from the paper not reporting it. For out-degree we were able to separate the two: the audit in §2.5 retrieved 26 full texts and recovered no additional out-degree, so the 22% reflects the literature rather than our coding depth. The same separation has not yet been done for persistence.

Keyword false positives. Two of the 30 quota-selected papers were false positives of the keyword screen, one from control theory and one from intra-agent decomposition. A screen that admits 7% noise at the selection stage is tolerable for a purposive sample and would not be for a systematic one.

Publication bias toward mitigation. The literature is dominated by papers proposing a method that improves a metric. Three of the four amplification findings in our set come from work whose primary contribution is a model or a diagnostic rather than a remedy [9, 13, 14]; the fourth is a mitigation paper that reports the failure it then fixes [34].

Purposive sampling. Restated because it governs every percentage in this section: these are properties of the coded set.

2.7 What would change these numbers

Three steps, in order of cost. A full-text pass over the 32 included studies aimed at the fields the §2.5 audit did not cover — persistence, and the channel-specific measures where a paper reports both prevalence and detectability. The out-degree question is settled, in the negative. Second, a coder independent of the first on a 20% sample, reporting κ for verification, topology and direction separately. Third, a pre-registered test of Propositions 1 and 2 on a corpus drawn by exhaustive screening of a fixed query set, at which point the proportions become estimates of something larger than themselves.

Until then the numbers in this section do what they can: they establish that the variables the framework needs are largely unreported, and that the one complete pattern in the data — the absence of any amplifying study with a verification channel — is worth an experiment.

3. A conditional account of error spread

This section supplies the vocabulary the rest of the paper uses. It has three jobs. It states what counts as a formalization, so that the six in §1 can be counted rather than asserted. It defines the quantities that make the sign of propagation decidable. And it shows how the six existing formalisms sit inside the resulting frame, which is what allows their disagreements to be read as differences in projection rather than as contradictions.

We build the account in the order a claim travels: the objects (§3.2), the one definition that does the most work (§3.3), the dynamics of acceptance (§3.4), the two ways a system can be harmed (§3.5), the operators that act at each stage (§3.6), and the two propositions that fix the sign (§3.7–3.8). §3.9 places the prior formalisms; §3.10 states where the account stops. Figure 1 shows the whole of it at once; the subsections that follow take it apart.

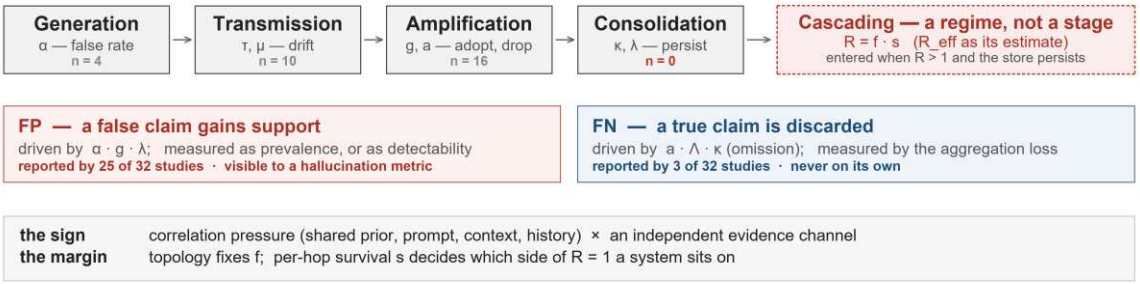


Figure 1. The account in one view. Four stages act on a claim's support; cascading is a regime entered when R > 1 and the store persists, not a fifth stage. The count under each stage is the number of the 32 included studies that address it: consolidation is addressed by none of them, because every memory system in the retrieval set was single-agent (§2.3). The two outcome channels are kept apart because a hallucination metric sees only the first; the FN channel is reported by three studies and never on its own.

3.1 What counts as a formalization

A review that claims to reconcile *six* formalizations owes a criterion for the number. We require three properties, and a paper must satisfy all three.

F1 — a state space. The work defines the object that varies across agents: a claim's truth status, a belief distribution, a context state, an influence weight.

F2 — a transition rule across an agent boundary. The work specifies how that object changes when it passes from one agent to another, or how it is aggregated across agents. A metric without an update rule does not qualify; a taxonomy without a state space does not qualify.

F3 — a measurable quantity. The work yields at least one parameter or threshold that can be estimated from observations and compared across systems: an escape probability, a reproduction number, an amplification factor, a divergence score.

The criterion is deliberately mechanical, because it is what makes the count auditable. Under it, six works qualify: the four-state Markov process [13], the graph network process [14], the claim-level cascade [12], the context-divergence model [15], the confidence-pooling model [16] and the cross-channel influence tensor [17]. Two adjacent works do not qualify, and we say so rather than stretching the definition: a topology-aware evaluation framework whose "topology" refers to a knowledge graph rather than an agent graph fails F1 and F2 [31], and a decomposition of debate failure into three pathways offers no state space [9]. Both are used as evidence in later sections; neither is counted.

Two of the six qualify only partially, and we mark them. The context-divergence model supplies a state metric and a synchronisation protocol, but its cross-boundary update is weaker than an explicit transition rule [15]. The confidence-pooling model has an explicit update equation, but its dynamical weight is a swept hyperparameter rather than an estimated quantity, so its F3 rests on derived indices [16]. The remaining four satisfy all three properties unambiguously.

3.2 Objects: claims, events, lineages

Definition 1 (claim). A claim c is a minimal, self-contained proposition that can be judged true or false. We take the claim, not the response, as the unit of analysis. A response can be right for the wrong reasons and wrong for the right ones, and both cases are invisible when only the final answer is scored [31].

Definition 2 (event). Running a multi-agent system produces events $v = (c, a, t, m)$, where agent a at time t disposes of claim c by mode $m \in \{assert, repeat, endorse, weaken, retract, ignore\}$. The mode matters: endorsing a claim and repeating it are different acts, and only the first adds support.

Definition 3 (lineage). A run induces a directed acyclic graph $G = (V, E)$ over events. An edge (u, v) exists when v's context contains u. $Anc(v)$ denotes the ancestor set of v.

The lineage is the object that the word "snowball" has been used to gesture at. On this definition it is not a metaphor: it is a graph, and the quantity that grows is support without evidence, accumulated along its edges.

3.3 Independence: the definition that does the work

The account turns on one distinction. Support for a claim can arrive as corroboration or as evidence, and the difference is provenance.

Definition 4 (independent evidence). An item x is independent of event v precisely when x's provenance is disjoint from v's ancestors:

$$\text{indep}(x, v) = 1 \iff \text{prov}(x) \cap \text{Anc}(v) = \emptyset$$

This is a strict test, and it is meant to be. An agent asked to review another agent's output has that output in its context. Its judgement is therefore a function of the claim's own lineage, and it counts as corroboration. The same holds when the reviewer is a different model: heterogeneity changes how the judgement is computed, not what it is conditioned on. Only an external tool return, a retrieval hit from a corpus the agents did not write, or a rule-based check over the claim's own surface satisfies Definition 4.

The test is severe enough that it reclassifies work that presents itself as verification. In one system, agents periodically exchange compressed state summaries and flag high-divergence conditions before joint reasoning, a procedure the authors describe as shared-state verification [15]. Under Definition 4 the exchanged summaries remain inside the agent set, so the procedure is a coordination protocol, not an evidence channel. This is not a quibble about naming. §3.7 predicts a different outcome for coordination than for verification, and the two are only distinguishable if the definition is applied.

Figure 2 makes the test concrete on a single claim. Two agents assert it, a memory record returns it with its provenance gone, and a tool returns it independently. The receiving agent counts four supporting items; the lineage supports three assertions and one item of evidence.

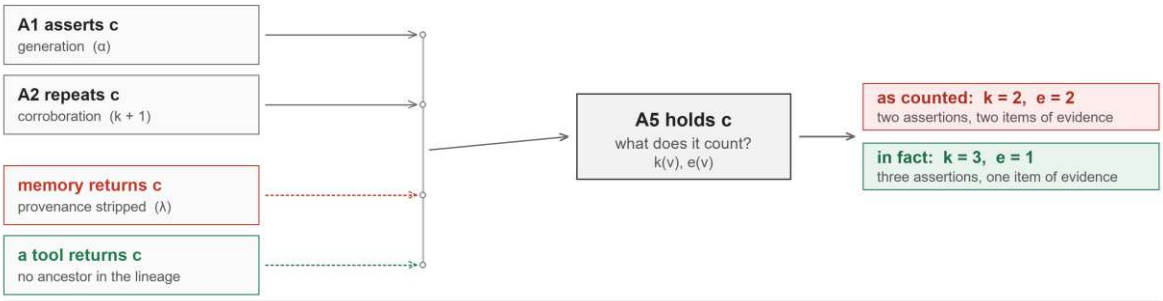


Figure 2. One claim with four incoming paths. Only the tool return satisfies Definition 4; the memory record fails it and is counted anyway. The receiving agent computes k = 2, e = 2, where the lineage supports k = 3, e = 1.

3.4 Two counts, and an asymmetry

For an event v concerning claim c, let $k(v)$ be the number of ancestors that assert c — the corroboration count — and $e(v)$ the number of independent items bearing on c. The pair (k, e) is the state of the claim at that point, and it is the object the six formalisms each describe a projection of.

Agents accept claims, and they also abandon them. These are not the same process, and treating them as one is the most consequential simplification in the existing literature. Write $\pi(v)$ for the probability that an agent accepts c, with prior log-odds θ_c from its parameters:

$$\operatorname{logit} \pi(v) = \theta_c + g \cdot k(v) + \beta \cdot e(v)$$

where $g \geq 0$ is a false-adoption coefficient and β carries the sign of the evidence. Separately, write $a \geq 0$ for the rate at which an agent holding a correct claim abandons it under social pressure. Most treatments assume $g = a$, that social influence is symmetric with respect to truth. Nothing requires this, and the difference is what determines which of the two harms in §3.5 dominates.

3.5 Two ways to be harmed

False-positive amplification (FP). A false claim gains support faster than evidence accumulates: $g \cdot \Delta k > |\beta \cdot \Delta e|$. This is the process the word "hallucination" usually names, and it is what an error-rate metric measures.

Truth suppression (FN). A correct claim loses support, even though no new false claim is introduced. Agents that hold the right answer defect under majority pressure, and an aggregation rule that returns the mode then discards it. The cleanest measurement is the aggregation loss:

$$\Lambda = \mathrm{P}(\text{correct} \in \text{candidate pool}) - \mathrm{P}(\text{aggregate} = \text{correct})$$

Λ is attractive because it needs no ground truth beyond the pool already generated. One study reports Λ up to 32.3 percentage points in homogeneous debate [9]; another reports a single scalar that separates omission bias from invention bias across 249,252 clause-level instances [35]. These are, to our knowledge, the only two operationalisations of FN in the multi-agent literature, and they were developed independently. What any rule can do with Λ is settled separately:

Proposition 3 in §5.2 shows that no aggregation rule removes it uniformly over pools, which is why Λ is a property of the pool rather than of the voting scheme.

The distinction has a consequence for evaluation. FP and FN are not symmetric with respect to what a hallucination metric sees. A system can report a falling hallucination rate while its aggregation loss rises, because the claims it discards were never generated as hallucinations — they were correct and then abandoned. An evaluation that scores only unsupported content is blind to the case where a correct claim is discarded in favour of an incorrect one, and a design that optimises against such an evaluation will drift into it.

3.6 Operators and a regime

We now assign the stages their operators. The assignment is not a claim of novelty at the level of mechanism — §3.9 shows that a recent detection framework proposes four mechanisms that map closely onto these — but it fixes which parameter each stage moves, which is what a governance analysis needs.

Table 4. Operators by stage.

Stage	Parameter	What it measures
Generation	α	P(false claim emitted | task, role)
Transmission	τ, μ	probability a claim's truth survives a hop; semantic drift per hop
Amplification	g, a	false-adoption and truth-abandonment per unit of corroboration
Consolidation	κ, λ	probability of persistence; rate at which provenance is lost on write
Cascading	R, Λ	branching ratio; aggregation loss

Two of these carry more weight than their brevity suggests.

Drift, μ. Transmission is not lossless copying. Claims are restated, compressed and specialised at each handoff, and the restatement can preserve fluency while detaching the claim from the evidence that supported it. One taxonomy of handoff failures separates data gaps from signal corruption from referential drift [36]; referential drift is μ. A protocol that replaces free-text messages with graph operations over a shared knowledge base targets this parameter directly [37].

Laundering, λ. When a claim is written to persistent memory, its provenance is typically not. What was one agent's conjecture returns on retrieval looking like a fact, and downstream agents count it as *e* rather than *k*. This is the mechanism by which cross-episode amplification can occur without any single episode showing growth: the write converts corroboration into apparent evidence. Work on memory poisoning describes the effect from the attack side [24]; work on memory consolidation describes it from the defence side, noting that once stored, an error becomes a persistent system-state failure [28]. It gives the consolidation stage its own operator rather than treating it as transmission that happens later.

3.7 Proposition 1: the correlation–verification condition

Proposition 1. Aggregation across agents reduces the error rate only if two conditions hold together: (i) the agents' errors are not strongly positively correlated, and (ii) at least one channel supplies evidence independent of the agent set in the sense of Definition 4.

The argument is the classical one about jury competence, applied to a case where its premise fails. Condorcet-type aggregation gains its power from independent voters; once errors are correlated, the effective number of voters collapses to the number of independent sources, not the number of instances. Multi-agent LLM systems violate (i) by construction: agents share pretraining, often share a prompt, share retrieved context, and — most consequentially — share the conversation history, so that a later agent's error is conditionally dependent on an earlier one. They violate (ii) whenever the verifier is another instance of the same model family reading the same context.

The two conditions partition the space in a way that predicts the sign.

Table 5. The correlation–verification condition.

	Independent verification present	**Absent**
High correlation pressure	convergence on truth	**FP amplification**
Low correlation pressure	self-correction	drift, without coherent growth

Neither condition alone is sufficient, and the asymmetry is testable. High pressure without verification produces amplification; absence of verification without pressure produces unstructured drift rather than a snowball, because there is no force pushing the population toward a common wrong answer.

The corpus supports the partition more cleanly than we expected. Among 24 included studies that make a directional claim, all nine that add an independent verification channel — retrieval, tool, rule-based or external knowledge base — report attenuation. Neither of the two studies that report amplification has one. The two features do not co-occur anywhere in the coded set. We did not construct this split; it is the distribution over the 24 directional rows. It should be read with the caveat in §2: the coding is abstract-level, and the split was re-tested against a second coder (§A.4).

3.8 The threshold

Let f be the fan-out: the number of agents or contexts that inherit a claim from a given event. Let s be the per-hop survival probability, $s = \sigma(\theta_c + g \cdot k + \beta \cdot e)$, which lies strictly inside (0,1). Define the branching ratio

$$R = f \cdot s$$

with the effective reproduction number R_{eff} as its empirical estimate.

Proposition 2a. Amplification compounds only if $R > 1$, which requires $f > 1/s$.

Proposition 2b. For a linear chain, $f = 1$, hence $R = s < 1$ for any survival probability. Extensive amplification is subcritical by construction.

Proposition 2b explains the first of the two studies that opened this paper. A chain cannot amplify the *number* of carriers of a claim, whatever the model, because each carrier has exactly one successor. The attenuation it measured is a topological consequence, not evidence that models correct themselves. The same reasoning explains why the second study amplifies despite also using a chain: it measures detectability, not carrier count, and detectability is governed by μ and λ rather than by R. Topology bounds one channel and leaves the other unconstrained.

Persistence acts on the horizon, not on the rate. A store that retains a claim extends the number of hops over which R compounds, from the length of a conversation to the lifetime of the store. The effect is multiplicative and one-sided. For $R > 1$, persistence turns a bounded perturbation into an unbounded one; for $R < 1$, it keeps a decaying error available for longer without making it grow. Memory therefore amplifies supercriticality rather than causing it. §4.4 treats consolidation as a distinct operator for a different reason — the type conversion it performs — and not because persistence by itself generates amplification.

Conjecture 1 (the threshold reading). If f ranges over roughly 2 to 9 in common configurations while s varies across models over a much wider interval, then s would decide which side of the threshold a system falls on, and f how far past it the system sits.

We offer this as a conjecture rather than a result, and the distinction matters. It follows from the ranges above only if those ranges hold, and no published study varies fan-out at fixed model, or model at fixed fan-out, while reporting both a reproduction number and the graph. What the conjecture would explain is the pattern of disagreement in §1. A model with low s would not cascade at any topology, which is consistent with the clinical study finding that topology was not reliably better [8]. A model with high s would cascade at almost any $f > 1$, which is consistent with the scale-free regime where cascading was strongest [14]. On this reading the two claims need not conflict; they may describe opposite sides of $R = 1$. The reading becomes testable as soon as one study reports R_{eff} across a fan-out sweep at fixed model, or across models at fixed fan-out — an experiment that, in our corpus, does not yet exist.

We are not the first to state the threshold. Reference [14] reports an effective reproduction number and shows a defence moving it from 1.08 to 0.81, and reports the strongest cascading in scale-free topologies. The contribution here is the separation of s from f and the reading of the dispute that follows from it, not the threshold itself.

Figure 3 sets the threshold beside the mechanism space that decides which side of it a system falls on.

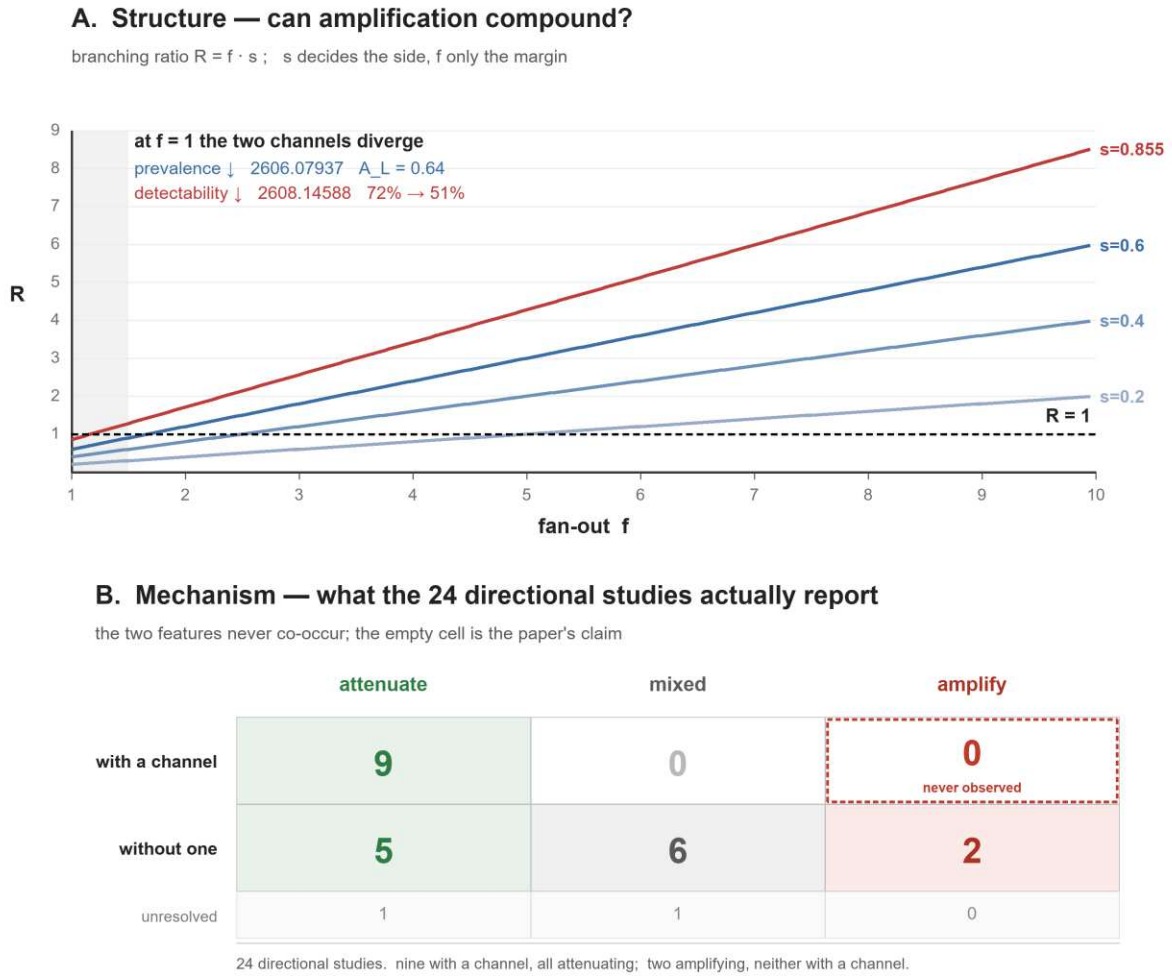


Figure 3. Structure and mechanism. (A) A chain is subcritical for every survival probability, so attenuation in prevalence is a topological consequence rather than evidence of self-correction; the one chain study that reports both directions does so because it measures detectability alongside prevalence. (B) The 24 directional studies in the coded set, by whether an independent verification channel is present and by the direction each reports. Nine studies have a channel and all nine attenuate; the two that report amplification have none. The cell marked zero is the paper's central empirical claim, and §A.3 states what would falsify it.

3.9 Where the six formalisms sit

With the vocabulary fixed, the six can be located. Each describes the same underlying *(k, e)* process and reports a different projection.

Table 6. The six formalisms and what each projection cannot see.

Work	Formal object	Quantity	Channel it sees	What it cannot see
Singh et al. [13]	first-order Markov over 4 states	escape probability	intensive (detectability)	topology; FP/FN split
Jamshidi et al. [14]	network process on directed graph	R_{eff}	extensive	laundering; detectability
Jamshidi et al. [12]	claim-level stochastic cascade	A_L, ρ, Γ	extensive	detectability; aggregation loss
Rodrigues et al. [15]	pairwise context divergence	CDS	transmission	direction of propagation
Han et al. [16]	confidence-normalized pooling	CI, TTC	amplification	persistence; recovery
Venkatesh et al. [17]	causal influence tensor	spectral amplification	extensive, multi-channel	FP/FN split; recovery

Read this way, the disagreements in §1 are mostly differences in projection. The two chain studies differ because they measure different rows. The topology-versus-model dispute would, on Conjecture 1, be a disagreement about which of *s* and *f* dominates, and therefore about where on the threshold the experiments sat. Absent that conjecture it stays unresolved, and we flag it as such.

What the table also shows is what none of the six supplies: a separation of FP from FN, and any treatment of recovery. Those gaps are the subject of §4 and §6. Table 7 lists the metric each formalism contributes, and makes the same point quantitatively: no metric in this literature is reported by more than one study.

3.10 Identifiability, and where this account stops

Three limits bound what the account can claim.

The coefficients are not yet estimable from published work. Almost no study varies *k* and *e* independently, so *g* and *a* cannot be separated from the reported aggregates; and only one paper reports an R_{eff} at all. The framework is therefore, at present, a scheme for organising and comparing evidence rather than a calibrated model. §6 proposes the minimal experimental design that would change this.

Conjecture 1 inherits the same limit and we do not want it read as settled. It rests on two studies, one on each side of the threshold, rather than on a sweep across it. A single counter-example — a fixed model that fails to cascade as fan-out rises past *1/s* — would retire it, and the governance advice in §5 would not survive intact if it fell.

Proposition 2b is a statement about expected carrier counts, not about every measurable quantity. A chain can attenuate in prevalence and amplify in undetectability at once, as the two opening studies jointly demonstrate. The proposition constrains the extensive channel only.

The strictness of Definition 4 has a cost. Some systems achieve correction through channels that are independent in a weaker, epistemic sense — a model's parametric knowledge of a fabricated entity, for instance, can catch a fabrication that no tool would flag [38]. Treating such channels as non-independent keeps the taxonomy clean and loses a real mechanism. We return to this in §6, where claim type is treated as a moderator of channel availability rather than as part of the channel definition.

Table 7. What each metric actually measures

The metrics in this literature are not competing estimates of one quantity. They are projections of the same process onto different observables, which is why studies using different metrics can reach opposite conclusions without either being wrong (§3.9).

Metric	What it measures	Channel	Extra annotation beyond claim truth	Reported by
Aggregate hallucination rate	share of unsupported content	FP	none	most of the corpus
Amplification factor A_L	ratio of final to initial claim-level score	FP, extensive	none	Jamshidi et al. [12]
Relative propagation ρ	per-hop change in claim score	FP, extensive	per-hop alignment	Jamshidi et al. [12]
Boundary escape probability	probability a claim changes state at a handoff	intensive	state labels per boundary	Singh et al. [13]
Effective reproduction number	carriers produced per carrier	FP, extensive	carrier definition	Jamshidi et al. [14]
Aggregation loss Λ	pool oracle accuracy minus aggregate accuracy	FN	none, if the pool is retained	Bertalanič et al. [9]
Risk Direction Index	omission bias against invention bias	both	typed claim categories	Yadav et al. [35]
Context divergence score	knowledge-state distance between agents	transmission	pairwise state probes	Rodrigues et al. [15]
Conformity index, time-to-consensus	agreement and convergence speed	amplification	none	Han et al. [16]
Spectral amplification	growth of propagation energy	extensive, multi-channel	per-turn trace, influence estimation	Venkatesh et al. [17]
Detection latency	turns from entry to flag	recovery	entry-point labels	proposed, §6.4

Three consequences follow from reading the table as a whole.

Only two rows have the property that makes a threshold claim testable: a quantity that can be estimated per system and compared across systems. The reproduction number is one. The aggregation loss is the other, and it measures the opposite channel.

Only one row describes the intensive channel in a way that could be compared across studies, and it is defined on a four-state scheme specific to one paper's pipeline. Everything else measures prevalence.

No row is reported by more than one study. The overlap column is empty for every metric, which is the quantitative form of the claim in §2.5 that this literature does not yet have commensurable quantities.

4. Stage by stage

4.1 Generation

Generation is the boundary condition of the whole account, and the literature on it is the most mature part of the subject. Single-agent hallucination has taxonomies [10, 11], benchmark suites, and detection methods several years old, of which consistency-based detection [39] is the one closest to the account here, since it asks whether a claim survives re-sampling rather than whether it is true. A recent survey organises the agent-specific case around the stages of one agent's workflow and enumerates eighteen triggering causes [19]. We do not re-survey that material. What matters here is smaller and less settled: the generation rate is not a scalar, and it is not the quantity that predicts what travels.

α depends on the claim type. The clearest measurement comes from legal workflows, where an aggregate hallucination rate of roughly 52% is reported across 249,252 clause-level instances, and where that average conceals a within-model gap of 38 to 40 percentage points between claim categories [35]. A system's headline hallucination rate is therefore an average over a distribution of claim types that differ in how they are generated and, as §4.2 argues, in how they survive a handoff. Two systems with the same aggregate rate can have entirely different propagation profiles.

α depends on the role. The same model placed in a planner, a critic and a summariser is emitting different kinds of claim under different constraints. Our corpus does not contain a clean measurement of this, and we flag it as a gap rather than assert a magnitude. What the corpus does show is that role assignment changes outcomes while the model is held fixed [16], which is consistent with a role-dependent α but does not isolate it.

Generation rate does not predict propagation rate, and the relation is not monotonic. Instructing a model to reason step by step raises the number of intermediate claims it emits, and each of those is a carrier that an answer-level metric never sees [40]. A fabricated entity is generated readily — most models tested produced a fictional Danish artist when asked to write about one [38] — and it is also easy to refute, because the claim is checkable against parametric knowledge. A subtle numeric substitution is generated less often and is much harder to catch once it has been restated as a derived figure [13]. The two properties that make a claim likely to appear and likely to survive are different properties, and a literature that optimises α treats them as one.

Two structural observations from evaluation work bear on this stage. Hallucinations concentrate at a small set of highly connected knowledge hubs, so generation is unevenly distributed over the concept space rather than spread across it [31]. And terminal accuracy decouples from trajectory

honesty: models reach correct conclusions through structurally flawed reasoning, which means that a low error rate at the output can coexist with a high error rate at the intermediate steps where claims are actually generated [31]. Auditing the trajectory rather than the answer is what makes that visible, and the audit frameworks built for it report failures that the answer-level metric scores as clean [41].

A generation-side intervention can create a failure rather than remove one. Enforcing structured reflection through constrained decoding does not improve self-correction in an 8B model; it replaces hallucination snowballing with a structural variant, in which the model justifies an early error inside the format it has been constrained to use [42]. The constraint relocates the error rather than stopping it. That is the difference between α as a rate and α as a location: a system can reduce the unstructured form of the failure while giving the structured form a place to live.

4.2 Transmission

Transmission is the stage the field names most often and measures least. Almost every paper in our corpus that discusses multi-agent failure mentions error propagation at inter-agent handoffs; six of the thirty-two report any transmission-level measurement, and two report a per-hop change in a claim-level quantity. The ratio is worth stating plainly, because a stage that is named in nearly every introduction and measured in six papers is not a stage with a settled account.

Four different things are called propagation at a handoff. The one systematic taxonomy we found separates them at the edge rather than at the agent: data gaps, where information needed downstream was never passed; signal corruption, where it was passed and degraded; referential drift, where the referent of a term shifts between agents; and capability gaps, where the receiving agent cannot act on what it received [36]. Only the third of these is drift in the sense of μ in §3.6. The other three are failures of the edge rather than of the message, and a metric that reports "propagation rate" without separating them cannot say which mechanism is operating.

Transmission re-types a claim rather than degrading it. The most detailed per-hop evidence comes from a four-stage pipeline that tracks injected hallucinations across boundaries and reports the state of the claim at each one: raw fact, derived computation, narrative prose, editorially approved conclusion, with per-boundary escape probabilities of 24.6%, 48.3% and 89.3% [13]. The probabilities are not uniform, and their pattern is the point. A claim does not fade as it travels; it changes kind, and each change strips the qualifications that made it defensible. A number that was true of a sample becomes a derived figure, then a sentence in a narrative, then a conclusion someone signed off on. Fidelity of text is the wrong measure for this, because the text is supposed to change.

What survives is the warrant, not the string. Read this way, τ in §3.6 should be defined over the claim's evidential warrant rather than over its wording. A handoff through a long context is not a neutral channel: models use their input unevenly, and material in the middle of a long prompt is retrieved less reliably than material at either end [43], so where a claim sits in the relay matters independently of what it says. The one study that tracks a claim-level quantity per hop finds the quantity declining across a three-agent chain [12], and the visual analogue of the same

process shows the mechanism: attention allocated to the evidence-bearing tokens falls across agent turns, so the relay loses the visual grounding while preserving the fluent description [34].

The edge is starting to be modelled directly. Two lines have moved from the agent to the interaction. One learns separate risk effects for correct and incorrect senders and composes them along an execution trace [44]; the other represents the run as a communication graph and estimates each step's reliability from local uncertainty plus uncertainty inherited upstream [45]. Both produce a per-edge quantity of the kind s denotes. Neither aggregates it into a reproduction number, so the local half of the measurement the framework needs is beginning to exist while the aggregate half still does not. That asymmetry is the same one §4.5 finds between attribution and reach.

The measurement gap has a specific shape. Both per-hop studies are chains, where "hop" is unambiguous. No study in our corpus measures a per-hop change in a graph, a tree or a shared-memory system, and the reason is structural rather than accidental: in a graph there is no single successor, so a per-hop rate is not defined without first fixing a path, and the field has not settled how to do that. Proposition 2 needs s, the per-hop survival probability, and the literature supplies it only for the topology in which it is easiest to define.

4.3 Amplification

Amplification is the stage at which a false claim's support grows faster than its evidence. In the notation of §3.4 it is the condition $g \cdot \Delta k > |\beta \cdot \Delta e|$: corroboration rises, independent evidence does not, and the acceptance log-odds move in favour of a claim that has earned nothing.

Two features make this the hardest stage to write about honestly. It has the most evidence — our retrieval returned 196 candidate records from the conformity and sycophancy queries alone — and the least agreement, because "debate" names a family of procedures whose members sit on opposite sides of that inequality [5, 6]. A review that reports "debate improves factuality" and "debate degrades factuality" as competing findings has not yet described the situation. The findings do not compete; they measure different pathways.

4.3.1 Three pathways rather than one

The clearest decomposition comes from a controlled study of homogeneous debate: ten agents, three rounds, three models in the 7–8B class, on two hard reasoning benchmarks, compared against isolated self-correction and a noise-injection control [9]. It separates three pathways and reports a magnitude for each.

Conformity. Agents adopt the majority answer without evaluating it, at a modal adoption rate of up to 85.5%. This is the pathway that moves g in §3.4: support rises because others assert, not because anything was checked. The single-agent form of the same failure is documented as sycophancy, where a model abandons a correct answer under social pressure [46]; debate supplies the pressure with a majority to converge on and a rationale to adopt. The rate is high at minimal peer exposure — a communication density of two neighbours already produces it — and rises with greater initial diversity among agents. That last result is counterintuitive and matters for design: heterogeneous starting points do not protect a debate, they give conformity more to work with.

The pathway does not need a round structure. In a single-turn design where one agent states a fact, a second defends a falsehood and the same model family judges, the forceful false agent overrides the truthful one at a rate that rises with its verbosity [47]. No peer exchange, no consensus, no aggregation rule — g moves because persistence is persuasive on its own. A system that treats multi-round debate as the risk surface has the mechanism in the wrong place.

Contextual fragility. Peer rationales destabilise reasoning that was previously correct, at a vulnerability rate of up to 70.0%. This is a transmission effect (μ in §3.6) rather than an amplification effect, and it is the one most easily missed: an agent can end the round with a worse justification while its stated confidence rises, because the new rationale is more fluent than the old one.

Aggregation loss. Plurality voting discards correct answers that were already in the pool, with an oracle gap of up to 32.3 percentage points. This pathway introduces no new error. It is the FN channel of §3.5, and it is orthogonal to both of the others: a debate can be free of conformity and still lose the correct answer at the vote.

The decomposition earns its place because the three pathways respond to different interventions. Conformity is reduced by cutting exposure or decorrelating agents; fragility by preserving the original rationale through the handoff; aggregation loss by keeping the candidate pool rather than committing to a mode. A single metric of debate quality averages all three and therefore cannot guide any of them.

4.3.2 The correlation is structural, not incidental

Proposition 1 requires that agent errors not be strongly correlated. In these systems they are correlated by construction, and it is worth being precise about why, because each source suggests a different remedy.

Agents share pretraining, so they share priors and share the misconceptions those priors encode. They frequently share the prompt, so they share the framing of the task. They share retrieved context when retrieval is centralised, and they share the conversation history whenever the debate is multi-round — which is the case that matters most, because shared history makes a later agent's error conditionally dependent on an earlier one. The dependence is not a bug in any component. It is what a shared context is.

Two measurements make the consequence concrete. Comparing an observed post-communication graph against a matched no-communication counterfactual shows that the same vote share can mean reliable agreement in one topology and over-confidence in another, because communication induces correlated failures that a vote count cannot see [48]. In multimodal settings the same point arrives as a requirement rather than a caveat: answer-level agreement is insufficient unless the image regions the agents relied on align, and a consensus built on unaligned evidence is agreement about different pictures [49]. Both results say that counting agreement measures the k term, and that the e term has to be checked separately or not at all.

The empirical signature appears in a study that varies only the aggregation structure at fixed agent count: seven agents arranged as star, hierarchical, ring or complete graph, with a confidence-normalised pooling rule and a swept self-weight [16]. Centralised structures decide quickly but

inherit the hub's competence and show same-model alignment bias. Distributed structures are more robust in the median and slower to converge. As connectivity rises, convergence accelerates and so does the frequency of what the authors call wrong-but-sure cascades: the group settles on an incorrect answer while its confidence climbs.

That last observation connects this stage to the intensive channel of §3.6. A wrong-but-sure cascade is amplification measured in confidence rather than in carrier count, and it is the same phenomenon that a four-state model observes as declining detectability [13] and that a spectral monitor observes as growing propagation energy [17]. Three formalisms, three instruments, one channel.

4.3.3 When the inequality reverses

The corpus contains a substantial set of cases where debate improves factual reliability, and they share a feature that is easy to state and easy to overlook: every one of them adds something that is not another agent's opinion.

Debate augmented with retrieval [50], a configuration in which each participant holds a distinct external tool [51], a fact-checking pipeline that interleaves claim detection and evidence retrieval with the debate rounds themselves [52], a medical debate in which the opposing agent retrieves contradictory images rather than merely arguing [53], a three-agent pipeline whose verifier is deliberately kept ignorant of the generator's reasoning [54], agents exchanging graph operations over a shared knowledge base in place of free text [37], and a generation pipeline with rule-based checks between stages [55] all report improvement. Read against §3.3, these are not variations on a theme of "more discussion". They are the subset that satisfies Definition 4, and Proposition 1 predicts their direction.

The same reading handles a result that looks like a counterexample. Giving agents a prior estimate of each peer's sycophancy, derived before discussion, reduces the influence of sycophancy-prone peers and improves final accuracy by 10.5 percentage points absolute [56]. Nothing external was added. What was added was information about the *reliability of the corroboration itself* — a discount on k rather than an increment to e. The inequality can be moved from either side.

Not every improvement in the corpus comes from an added channel, and the exceptions sharpen the claim rather than weakening it. Four designs report attenuation with no independent evidence at handoff: a diagnostic reframing in which agents play undercover to expose which of them fabricated [57], a mediator model that reweights peers by their estimated uncertainty while remaining an LLM reading the same claims [58], an introspection scheme whose cross-modal perspectives are other agents rather than retrieval [59], and a debate that presets opposing stances to override a model's attachment to its first answer [60]. Under Proposition 1 these belong in the low-correlation cell of Figure 3B, not the channel cell: what they act on is the dependence among agents rather than the supply of evidence. Read this way the corpus supports a one-sided statement — a channel is sufficient for attenuation here, and not necessary for it — which is exactly the asymmetry §2.5 reports. A symmetric reading, that channels are what produce attenuation, is not supported by these four rows, and we do not make it.

4.3.4 What this stage's evidence does not establish

Three cautions bound the section.

Most of the amplification evidence is about error, not about hallucination. The controlled debate studies score task accuracy on reasoning benchmarks; a wrong answer there is an error, and only some errors are hallucinations in the sense of unsupported or fabricated content. The two literatures overlap and are not the same, and we have not found a study that measures the conformity pathway on a hallucination metric.

The direction labels in our corpus conflate the pathways. No included study reports both a prevalence measure and a detectability measure, so a paper's headline — "debate helps", "debate hurts" — cannot say which of the three pathways moved. This is why §2.5 reports the split as an absence of co-occurrence — nine channelled studies, all attenuating; two amplifying studies, neither with a channel — rather than as a causal claim. The coding records what each paper concluded, and the pathways are our reading of why.

The signals a debate system would use to police itself are degraded by the failure it is policing. Log-likelihoods and self-reported confidence are the standard priors for pruning or reweighting peers, and they become unreliable under hallucination, which is why a posterior-guided scheme has to treat them as priors to be revised by peer challenge rather than as measurements [61]. Corroboration cannot be discounted by a quantity that the hallucination has already corrupted.

The magnitude of g is unestimated. Every study above varies the procedure and reports an outcome; none varies corroboration and evidence independently, which is what separating g from β would require. Until that experiment exists, the coefficient in §3.4 is a way of stating the hypothesis, not a measured quantity.

4.4 Consolidation

Two of the five stages in this paper have substantial literatures. Consolidation has none in the setting we care about, and the reason is not that the mechanism is unimportant. It is that memory research and multi-agent research have only just met, and not yet on the question this paper asks. The single-agent side is mature: memory streams that summarise and retrieve prior experience [62], paged memory hierarchies that manage a fixed context window [63], verbal self-reflection written back into the agent's own context [64], and skill libraries that accumulate executable procedures across episodes [65]. Every one of those designs assumes a single reader of the store. Seven memory systems entered our retrieval and all seven were excluded on the same ground: one agent, no inter-agent channel. What follows is therefore an argument about a stage rather than a survey of one, and we mark it as such.

A memory write is not a delayed message. If it were, consolidation would be transmission with a longer gap and would need no separate operator. The operation that distinguishes it is a change in evidence type. When Agent A asserts a claim, the claim arrives downstream as corroboration: its provenance is visible, and an agent that weighs it knows it is weighing an assertion. When that same claim is written to memory and later retrieved, the write typically does not carry the provenance. It returns as a record. Downstream agents count it as evidence in the sense of §3.3,

and the transition from *k* to *e* is what λ in §3.6 measures. This is the only mechanism in the framework that converts a source type rather than moving or weighting one.

The consolidator is itself a generator. A memory system does not copy; it extracts, summarises and rewrites, and each of those operations re-authors the content. The point is not incidental to security. Work on memory poisoning reports that attacks which assume injected content is stored verbatim fail against modern pipelines, because selective extraction and rewriting intervene, and that the successful attacks are those engineered to survive those stages — binding a trigger to its payload so they are co-extracted, and shaping triggers to mimic named entities so that rewriting preserves them [24]. A consolidator that rewrites has an α of its own. Its output is a generation, and it is a generation that arrives without the markers of one.

Three failure modes, mapping onto the two channels. Work on trustworthy consolidation separates what can go wrong into omission of important information, corruption of existing memory, and introduction of unsupported content [28]. These are not three flavours of the same error. Omission is the FN channel of §3.5 — the correct fact was there and the write lost it. Introduction is the FP channel. Corruption is drift, and belongs to transmission. A memory system can therefore damage a system's factual state along both outcome channels at once, which is why the stage needs its own treatment rather than being folded into either.

The ratchet. The asymmetry that makes consolidation a distinct design problem is between writing and un-writing. A recent memory layer states the situation directly: existing agent memory systems are built around forward-only evolution, accumulating and overwriting knowledge with no principled mechanism to inspect, version or revert prior states, which leaves agents brittle under correction and memory corruption *after* they have already been exposed to subsequent information [30]. That is the ratchet described in §3.6, in the words of a system built to defeat it. Writes are cheap, cumulative and persistent; retractions are expensive, lossy, and in most deployed designs unavailable. Persistence alone does not amplify — the note in §3.8 shows it is a multiplier on supercriticality rather than a cause — but persistence combined with laundering turns a transient error into a durable premise, and the durability is what makes the downstream branching ratio matter over time rather than over a turn.

What would make this a multi-agent problem. Two conditions have to hold together. The store must be reachable by agents other than the one that wrote to it, whether by sharing or by episode boundary; and the write must not preserve provenance. Under those two conditions the laundering argument applies and λ is non-zero. Where memory is per-agent and ephemeral, consolidation is caching and carries none of this. Our corpus contains no system that satisfies both conditions *and* measures the resulting parameter, which is the honest statement of the gap: the mechanism is documented on one side and the setting exists on the other, and no study joins them.

The setting now exists, which sharpens the gap rather than closing it. A modular memory system couples six memory types — core, episodic, semantic, procedural, resource and a knowledge vault — to a multi-agent framework, so the store is reachable by agents other than the one that wrote to it [66]. It falls outside our inclusion rule because it studies memory quality rather than hallucination, not because it is single-agent, and that distinction matters: the multi-agent store is no longer hypothetical. Architectural work now frames multi-agent memory as a systems problem and names consistency across agents as the most pressing open one, separating

shared from distributed stores precisely because the write path crosses an agent boundary [67]. Benchmarks evaluate memory along in-episode and cross-episode scope [68], and surveys organise the mechanisms by storage, reflection and abstraction [69]. What none of them reports is the quantity λ denotes: whether a claim written to a shared store comes back with its provenance intact. The stage has systems, a named problem and a benchmark. It does not yet have the measurement.

4.5 Cascading

The first four stages act on a claim's support. Cascading acts on its scope. The distinction is the reason this subsection is about a regime rather than a stage: amplification changes how strongly a claim is held within a run, and cascading is what happens when the failure leaves the run — into other tasks, other episodes, other users, or the deployed system. The threshold is the one from §3.8, $R > 1$, and the scope crossing requires the persistence discussed in §4.4. Neither condition alone produces a cascade: a supercritical process confined to a single conversation ends when the conversation does, and a persistent store holding a subcritical error accumulates debris rather than propagating it.

The regime has been measured once, and the measurement is marginal. A six-agent heterogeneous system evaluated across five recursive reasoning rounds and several communication topologies reports an effective reproduction number of 1.08 without defence, falling to 0.81 with a defence that adds external claim verification and selective isolation [14]. Read against §3.8, the undefended system is barely supercritical and the defended one is comfortably subcritical, so a modest intervention moved the system across the threshold in the intended direction. That is the whole of the direct evidence. One number, one system, one margin. Every other use of the word "cascading" in this literature is a qualitative description or a post-hoc reconstruction.

Three regimes, not two. A branching account offers cascade ($R > 1$) and decay ($R < 1$). Failure analysis names a third: errors that persist inside unresolved interaction loops without spreading, which one attribution framework treats as a mechanism distinct from cascading propagation [70]. The deliberation benchmark offers an instance — of eighteen runs of a twelve-agent jury, seventeen ended without a verdict [71] — where the system neither converged nor dispersed. Whether a loop is $R \approx 1$ in the branching sense is not established, and we do not claim it. The point is narrower: a framework with two regimes will classify a stalled system as one of them incorrectly, and stalls are common enough in this literature to be worth a name.

R may be a property of the system and the task together. The single clearest counterexample to a system-level reading comes from a controlled comparison of synchronisation strategies. Naive full-broadcast synchronisation raises the hallucination rate by 34% over no synchronisation, with a large effect size, in multi-agent travel planning; the same comparison in software project planning shows every condition converging below a hallucination rate of 0.2 [15]. The mechanism the authors propose is specific: the effect depends on whether a single erroneous shared belief is reused across evaluation dimensions, which is true of itinerary planning and not of a codebase. If that reading holds, R is not a constant of an architecture but a function of architecture and task structure, and the framework in §3.8 would need a task term. We flag this as

an open problem rather than absorb it, because one comparison across two domains cannot distinguish a task effect from a domain effect from a sample-size effect.

A limit of the branching account. Proposition 2 uses f as a mean fan-out. The strongest cascading in the one measured system appears in scale-free topologies, with a reproduction number of 1.21 against 1.08 in the aggregate [14]. A scale-free graph does not have a representative fan-out: most agents have few successors and a few have very many, and a cascade is driven by the tail rather than the mean. A mean-f threshold therefore understates the risk in exactly the topologies where cascading is most likely, and the correct object is the out-degree distribution. We state this as a limitation of the formal account in §3.8, not of the empirical finding.

The topology evidence is not monotone in sparsity either, and the shape it takes is the one the out-degree distribution implies. A causal analysis that varies connectivity finds moderately sparse topologies achieving the best trade-off: dense graphs propagate errors efficiently, and very sparse ones lose the diffusion a task depends on, so the interior is where propagation is suppressed without the information flow collapsing [72]. That is the f term and the second outcome channel moving in opposite directions as connectivity changes — fewer edges lower R and raise the chance that a correct claim never reaches the agent that needed it.

Attribution is not measurement. Four papers in our corpus reconstruct where a failure began: by tracing information dependencies across agents [73], by ranking candidate root causes from execution logs [74], by annotating a benchmark of failed trajectories [75], and by self-improving semantic attribution [76]. A fifth, outside the coded set, models cascading propagation alongside persistent interaction loops in order to identify a decisive agent-step pair [70]. A taxonomy of 1,600 annotated traces across seven frameworks separates fourteen failure modes into system design, inter-agent misalignment and task verification, with inter-annotator agreement of 0.88 [23]. All of this answers *where a failure started*. None of it answers *how far it travelled*, which is the question the regime poses. The asymmetry is worth naming because it explains why the field can have a mature vocabulary for root causes and a single measurement of spread.

5. Governance

5.1 The matrix

The four stages and the two outcome channels give a two-dimensional space in which an intervention can be placed: what it acts on, and which harm it moves. Table 8 fills it. The rightmost column is the part that matters, because the literature is very unevenly distributed across the rows.

Table 8. Interventions, the parameters they move, and their effect on the two outcome channels. ↓ lowers the channel and ↑ raises it. — means the intervention does not act on that channel by construction; ? means the account makes no prediction for it.

Intervention	Parameter	FP	FN	Corpus support
Add a channel independent per §3.3	e, β	↓	↓	nine studies, all report attenuation
Isolate first, then aggregate	g	↓	↓	one study; isolated arm beats debate
Preserve the full candidate pool	Λ	—	↓	derived from the oracle gap; untested
Cap fan-out, or break the chain	f	↓	↑	derived from $R = f \cdot s$ and §4.5; untested
Require dual-source writes, keep provenance	λ	↓	—	mechanism documented; parameter unmeasured
Fewer rounds, forced convergence	g	↓	↑	supported; the oracle gap is the mechanism
Use decorrelated models rather than role prompts	g	↓	↓	derived from Proposition 1; untested
Increase initial answer diversity	g	↑	?	one controlled study: conformity rises with diversity
Raise the abstention rate	—	↓	↑	no corpus evidence

The two heterogeneity rows are kept apart because they are not the same intervention and they point in opposite directions. Drawing agents from different models lowers the shared prior that Proposition 1 makes responsible for correlated error, so both channels fall. Giving agents of the *same* model different starting answers does the opposite: a controlled study finds that conformity rises with initial diversity, because disagreement supplies the majority something to converge on [9]. A design that reads "increase diversity" as one instruction will get whichever of the two it implements, and the table previously collapsed them into a single row marked with a question mark. We have separated them rather than leave the sign undetermined. §6.2 treats the diversity manipulation as the cheapest available handle on g and states what each sign would imply for the account.

Four of the nine rows have direct support, and for three of those the evidence is a single study. Three are predictions of the formal account with no experiment behind them, and one has a documented mechanism but no measured parameter; we mark all four as such rather than presenting the table as a validated instrument. The last row has no evidence at all and is included because it is a design option that teams reach for, and because its effect on the second channel is not what its advocates assume: an abstention converts a wrong answer into a missing one, which

lowers the measured false-positive rate and raises the number of questions the system does not answer.

5.2 A vote cannot create information

The rows of Table 8 split into two classes, and the split is not a matter of degree.

Proposition 3 (the aggregation ceiling). Let a run produce a candidate pool P, a finite multiset of answers drawn from a set A, and let $c* \in A$ be the correct answer. An aggregation rule is any function ρ: $P \to A$, possibly randomised, that returns one element of the pool. Write $u = \Pr(c* \notin P)$ for the probability that the pool contains no correct answer, and $\Lambda(\rho) = \Pr(c* \in P \text{ and } \rho(P) \neq c*)$ for the aggregation loss of §3.5. Then:

- (i) $\mathrm{err}(\rho) = u + \Lambda(\rho) \geq u$ for every rule ρ;

- (ii) no deterministic rule is correct on every pool that contains the correct answer;

- (iii) therefore no intervention that leaves the distribution of P unchanged can move the floor u, and no rule choice removes the aggregation loss uniformly over pools.

Proof. For (i), the events $\{\rho(P) \neq c*\}$ and $\{c* \notin P\}$ are nested, and $\{\rho(P) \neq c*\}$ is the disjoint union of $\{c* \notin P\}$ and $\{c* \in P, \rho(P) \neq c*\}$, of probabilities u and $\Lambda(\rho)$. For (ii), take $A = \{a, b\}$ with $c* = a$ and consider $P_1 = \{a, b, b\}$ and $P_2 = \{a, a, b\}$. A rule that returns the plurality is right on P_2 and wrong on P_1; a rule that returns the minority is right on P_1 and wrong on P_2. A deterministic rule computes a single function of the multiset, so it cannot be right on both, and the two pools present the same 2:1 split. For (iii), u is a function of the distribution of P alone, and P is formed before any rule is applied; by (ii) $\Lambda(\rho) > 0$ on some pool for every ρ. ■

Clause (ii) is the load-bearing one, and it is worth restating without symbols. A two-to-one majority in favour of a wrong answer and a two-to-one majority in favour of the right one are indistinguishable to any rule that only counts. No weighting, no convergence criterion and no number of rounds separates them, because the information that separates them is not in the vote. That is the precise sense in which aggregation-side interventions are zero-sum between the two channels: they select different points on a frontier the pool fixes, and the frontier does not contain the ideal point unless $u = 0$.

What the proposition does not say matters as much. It does not say aggregation is useless. On the pools a system actually encounters, a well-chosen rule may recover most of Λ, and the measured cases below are improvements of exactly that kind. It says that the improvement is not available by design alone, so a team that can only change the rule is selecting a point on a curve it does not control.

The measured cases behave this way. Forced convergence reduces the influence of a wrong plurality and discards correct answers already in the pool, with an oracle gap of up to 32.3 percentage points [9]. Removing consensus altogether avoids the collapse and pays for it in a different currency — a consensus-free debate still has to select an answer [77]. Discounting untrustworthy peers by their sycophancy ranking improves final accuracy by 10.5 points absolute [56], and it does so by reweighting corroboration, which is an aggregation-side move: the improvement is bounded by what the pool contained.

An intervention that adds a channel independent in the sense of Definition 4 is not subject to this ceiling. It changes the pool rather than the weighting, so it can raise oracle accuracy instead of merely recovering it, and it lowers both channels at once because a claim that fails independent check is removed before it can acquire support while a claim that passes acquires support it would not otherwise have had. Retrieval is the standard instance [78], and the verification methods built on it — decomposing a claim into checkable questions and answering them against retrieved sources [79], or retrieving field-level evidence for each citation rather than judging the citation as a whole [80] — are the part of that literature that satisfies Definition 4. This is the formal content of the claim that verification dominates reweighting. All nine studies in our corpus that add such a channel report attenuation, and neither study that reports amplification has one.

Adding a channel is not the only pool-changing move, and the table's second row is the other one. Isolating the agents before they aggregate adds no external information and no verification of any kind. What it changes is the *distribution* of claims that enter the pool: independent judgments disagree, so the pool contains more distinct candidates and its oracle accuracy rises. Aggregation over that pool can then recover more than aggregation over a pool formed by agents who had already read each other. The dividing line is therefore not "external versus internal" but whether the intervention can raise the pool's ceiling. Isolation and verification both can. Reweighting cannot.

The distinction has a practical reading. Aggregation-side interventions are cheap, require no external access, and are the ones the literature mostly proposes. Their benefit is capped by the quality of the pool, and their cost is paid in the other channel. Evidence-side interventions cost an external dependency and buy a change in the pool. A team choosing between them is choosing between recovering part of a gap and closing it.

Figure 4 sets the two outcome channels against the one intervention in our corpus that has been observed crossing the threshold.

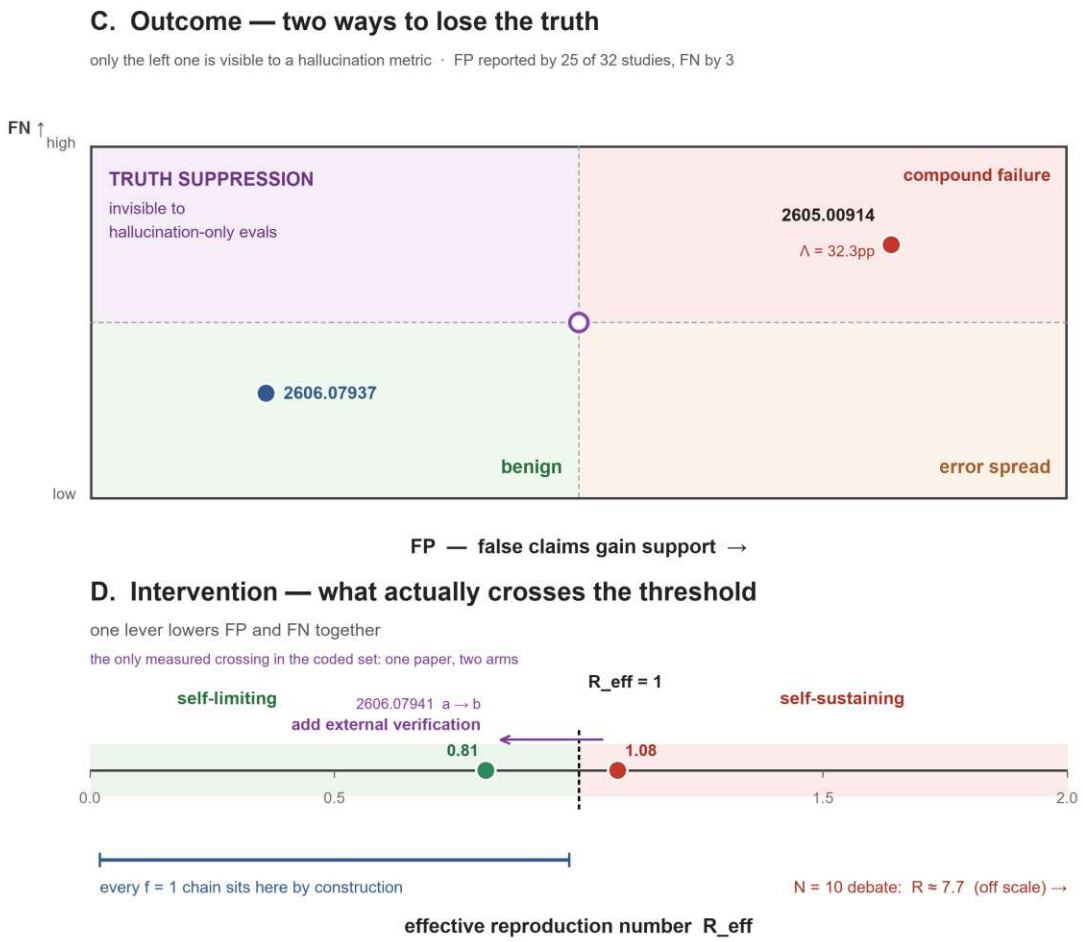


Figure 4. Outcome and intervention. (C) Two harms, only one of which a hallucination metric sees. The false-positive channel is reported by 25 of the 32 included studies; the truth-suppression channel by 3, and never without the first. (D) Adding an independent channel moves the reproduction number from 1.08 to 0.81 across the threshold, while every chain lies below it by construction and a ten-agent homogeneous debate lies far above it. This is the only crossing measured anywhere in the coded set, and it comes from one paper with two arms.

5.3 Recovery as a stage with its own metrics

The literature treats repair as a subcategory of mitigation, and the conflation hides a difference in kind. Prevention reduces the rate at which false claims enter a system; recovery determines how long a false claim that has entered survives. These have different costs, different failure modes, and different measurability. A prevention mechanism that works cannot be observed doing so, because the errors it removed never happened. A recovery mechanism that works can be timed.

We separate six actions and give each a parameter it moves or a quantity it is measured by.

Prevention reduces α, and inherits the measurability problem above.

Detection is measured by latency — the number of turns or tokens between a claim's entry and its being flagged. Latency is observable without counterfactual reasoning about what would

otherwise have occurred, which makes it the most tractable recovery metric in the set. Detection presupposes attribution, and attribution is a separate problem: identifying which agent introduced an error and at which step is what makes a latency measurement possible at all, and it has been approached as its own diagnosis task rather than as a by-product of scoring the final answer [76]. The deployed form is a runtime monitor that revises an answer after reasoning has been exchanged, and what such monitors report is an asymmetry rather than a net gain: some initially correct agents are misled by their peers even as others are corrected [81]. A detection rate that counts only the recoveries would read as a success. Locating the evidence that would have prevented a claim is a further step, and multimodal debate systems now report the responsible image region alongside the correction [82]; that is attribution at the level of the source, not just of the agent.

Containment reduces R by limiting the number of agents or tasks a flagged claim can reach: fan-out caps, circuit breakers, isolation of a suspect agent. The one measured defence combines containment with external verification and selective isolation, and moves R_{eff} from 1.08 to 0.81 [14].

Rollback targets κ^{-1}, and is the action with the weakest support in deployed systems. A memory layer built for it states the problem directly: existing systems evolve forward only, with no principled mechanism to inspect, version or revert prior states [30].

Inoculation raises the fraction of agents that will not accept an unverified claim, and is the recovery action closest to a design pattern rather than a mechanism. Giving agents a prior estimate of each peer's sycophancy is an instance [56].

Degradation converts an unresolved claim into an explicit absence: abstention, escalation to a human, or a flagged low-confidence output. It is the only action that reduces harm by reducing coverage, which is why its effect on the second channel must be reported rather than assumed. Clinical deployments are where the trade-off is least avoidable, and the guardrail work in that setting reports the same tension in its own vocabulary: the safe response and the helpful response diverge, and the system has to choose which to fail on [83]. Auditing after the fact is the complementary move — instead of gating the answer, test it against an adversarial set built from the failure the deployment most fears, which for regulatory knowledge means questions whose historically correct answers are now withdrawn [84]. Auditing buys the coverage back but pays in latency, since it runs after the claim has already been emitted.

Uncertainty is becoming the trigger for this set. Four systems now use a per-step reliability estimate to decide whether to verify, re-run, isolate or stop: inherited uncertainty propagated along the communication graph [45], an uncertainty signal injected into role-based development so that a low-confidence intermediate output is not passed downstream [85], semantic entropy and density used to choose between verification and re-sampling [86], and verbalised confidence written into memory so that it constrains later decisions [87]. Read against Table 8 these are not a new row. They are triggers for containment and degradation, and their value depends on a question §4.3.4 leaves open: whether the signal they act on survives the failure it is meant to detect. A monitor whose confidence is inflated by the hallucination it is monitoring will decline to intervene exactly when intervention is needed.

One design result from this set is worth isolating because it is cheap and general. The most effective verifier is not the most informed one. A pipeline that deliberately withholds the generator's reasoning from its verifier outperforms one whose verifier sees the full trace, because a verifier that reads the reasoning inherits its errors — the confirmation-bias failure that §3.3's provenance test predicts [54]. The same weakness is visible wherever a model is asked to judge text produced in its own idiom: judge models show measurable position, verbosity and self-preference biases, and they favour their own outputs at rates that make an unblinded judge a poor verifier [88, 89]. Information asymmetry inside a pipeline is a way of manufacturing the independence that Definition 4 requires, and it costs nothing external.

5.4 The ratchet as a bound

The argument for treating recovery as first-class does not need a normative premise. It follows from an asymmetry in cost.

Suppose writes to persistent state succeed with probability κ and are effectively irreversible, so $\kappa^{-1} \approx 0$ for most deployed systems. The expected number of durable false claims in the store then grows as $\alpha \cdot \kappa \cdot w \cdot T$, where w is the write rate and T the operating time. Prevention reduces the slope of that line. Nothing in the prevention set reduces its intercept, and no amount of prevention makes the accumulation bounded. Only an action that removes stored claims — rollback, or a verifier that runs before the write rather than after — bounds the level.

This is why the ordering in Table 8 matters more than the individual entries. A system with an excellent verifier and no rollback has a store that accumulates whatever the verifier misses, and the misses are precisely the claims that were difficult to check. A system with a mediocre verifier and a working rollback can recover from its misses. The two designs differ in the exponent of their long-run error, not in their steady-state error rate, and steady-state error rates are what the literature reports.

The asymmetry also explains why memory poisoning is a productive attack surface out of proportion to its success rate. An attack that plants one claim permanently, and that is engineered to survive the write path's extraction and rewriting, obtains a durable premise for the cost of a single interaction [24]. Against a system with $\kappa^{-1} \approx 0$, the exchange rate favours the attacker by construction.

6. Agenda

6.1 The experiment that would settle the central pattern

The most consequential finding in this paper is the pattern in §2.5: all nine studies with an independent verification channel report attenuation, neither study reporting amplification has one, and the two features do not co-occur anywhere in the coded set. We gave three readings for it and could not choose between them. They are separable by one design.

The design is a factorial with fan-out and verification as factors, holding model, task and claim set fixed. Three levels of fan-out — a linear chain, a two-way fan-out, and a four-way fan-out — crossed with two verification conditions, absent and retrieval-based, with injected false claims of

known ground truth and a claim-level tracker maintained across rounds. Primary outcome: the effective reproduction number estimated from the growth or decay of the set of carriers. Secondary: aggregation loss, and detection latency.

Three readings make three different predictions about this design.

If the association is **causal**, adding the verification arm should lower R_{eff} at every level of fan-out, and the fan-out main effect should be visible within the no-verification arm. The interaction is the interesting term: verification should matter most at high fan-out, because that is where corroboration accumulates fastest.

If the association is a **research-fashion artefact**, the effect should appear only in comparisons across studies, not within one. A single team running both arms with the same instrumentation would find no difference, and the corpus-level pattern would be an artefact of who chooses to build verifiers.

If the association reflects a **division of labour**, verification should lower R_{eff} but the studies that measure R_{eff} should be exactly the ones that never build one — which is what our corpus shows, and which means the effect would be real and unobserved at the same time. This reading predicts that the factorial will work and that nobody will run it.

We cannot distinguish these from the literature. One team running the factorial distinguishes all three.

6.2 Two coefficients that are not yet separable

Proposition 1 rests on g and a being distinct: false-adoption and truth-abandonment respond to the same variable, corroboration, but need not respond equally. Every study we coded varies the procedure and reports an aggregate outcome, which confounds the two. A design that separates them holds the claim pool fixed and varies the number of agents asserting a claim (k) independently of the amount of independent evidence (e), recording adoption and abandonment as separate outcomes rather than as a net accuracy.

The measurement is a pair of marginal effects: the change in acceptance log-odds per unit of corroboration, and the change in abandonment probability per unit of corroboration. If the two are equal, the symmetric assumption that most of the literature makes implicitly is correct and §3.4's asymmetry is unnecessary. If abandonment exceeds adoption, the mechanism behind the clinical result — that collaboration redistributes what models can already do rather than supplying what they cannot [8] — has a specific form, and the FN channel dominates at high correlation.

This is the cheapest of the four experiments and the one with the highest chance of a negative result. It should be run first.

A handle on g alone is already available, and §5.1 records what it did. Raising the initial diversity of agent answers moves g without touching e, and in the one controlled study that varies it the effect runs the wrong way for the design intuition: conformity rises when agents start from different positions, because disagreement gives the majority something to converge on [9]. A group that increases diversity to guard against groupthink is therefore moving the false-adoption coefficient in the direction it did not intend. That makes the manipulation a test of this account

rather than only a measurement. If the sign holds with k and e held fixed, g is a function of the spread of the pool and has to be modelled as one; if it reverses at greater model heterogeneity, the two rows of Table 8 are the reason, and the pathway decomposition in §4.3.1 is incomplete. Either outcome is informative, which is what makes this worth running early rather than last.

6.3 The operator nobody has measured

λ, the rate at which provenance is lost on a memory write, has no measurement in the multi-agent literature. The mechanism is documented from two directions: attacks engineered to survive extraction and rewriting [24], and defences that treat a stored error as a persistent system-state failure [28]. Neither measures the conversion rate.

The measurement is direct. Give a claim a known provenance — one agent's assertion, with a record — write it to shared memory, retrieve it after a fixed interval, and ask a fresh agent whether the retrieved item is supported. λ is the excess probability of a positive answer over a control in which the same content is presented with its provenance intact. The prediction from §4.4 is that λ rises with the number of consolidation cycles, because each rewrite strips more of the surrounding context that carried the provenance.

This experiment has a property the others lack: it needs no ground truth about the claim, only about its provenance. That makes it cheap, and it makes it the natural first target for a group without an annotation budget.

6.4 Recovery, with numbers attached

§5.3 separated six recovery actions and gave each a parameter. Only one of them has a natural primary metric, and it is the one the literature does not report: **detection latency**, the number of turns or tokens between a claim's entry and its being flagged. Latency is measurable without counterfactual reasoning, it is comparable across systems, and it is the quantity a deployed operator would actually care about.

A recovery benchmark built on latency would need three things the current benchmarks do not supply: claims with known entry points, a flagging mechanism whose false-positive rate is reported alongside the latency, and a rollback condition so that the effect of repair can be separated from the effect of prevention. The last is what would test the accumulation argument in §5.4, which currently rests on an illustrative expression rather than a measurement.

One form of evidence is available now and is not being collected: longitudinal incident records from systems in continuous operation. A single production agent runtime has been documented over eight weeks with full root-cause postmortems, yielding a mechanism-oriented taxonomy of silent failures — the class in which the error signal never reaches a human in actionable form [90]. That is the natural venue for latency and rollback claims, because a long-running system is the only place where durability and recovery can be observed rather than assumed, and because postmortems record the timing that a benchmark would otherwise have to reconstruct.

6.5 A reporting standard

Four quantities would make this literature commensurable, and none of them requires new science — only that authors report what they already know.

The carrier set. A reproduction number is meaningless until it is stated what counts as a carrier. An agent that re-asserts a claim is a carrier; whether a memory write is one is a modelling decision that changes f from 1 to 2 in a chain and therefore changes whether the chain can cascade at all. This definition should precede any cascade claim.

Aggregation loss. Λ is computable from a single run whenever the candidate pool is retained, needs no additional annotation, and is the only FN measure in the literature that has this property. One study reports it [9]. It should be a default.

The out-degree distribution. Twenty-two percent of our coded set reports enough structure to recover a mean out-degree, and §4.5 shows that the mean is the wrong summary in exactly the topologies where cascading is most likely. Reporting the distribution costs a sentence.

The comparator. "Amplifies" and "attenuates" are relative terms. Two studies can both be correct and appear to disagree because one compares against the first hop and the other against a single-agent baseline.

6.6 One question we cannot close

Whether R is a property of a system or of a system and a task together remains open. The travel-versus-software contrast in §4.5 is the only evidence, and it is two domains. If R is task-moderated, then every claim in this paper about thresholds is conditional on a task distribution that nobody has characterised, and the governance advice in §5 inherits that condition. We would rather state the dependence than absorb it into the framework, because a framework that claims to cover a moderator it has not measured is harder to correct later.

7. Conclusion

Two studies opened this paper. Both tracked a hallucinated claim through a linear pipeline, and both reported a headline pointing the other way. The account built here says they are both right, and says why: one measured how much of the text lacked support, the other measured whether a reader could still detect the lack. A claim can become rarer and harder to see at the same time. A literature that reports one of these quantities at a time will keep producing such pairs, and the pairs will keep reading as disagreements.

For a reader facing a new system, the framework's contribution is an order of questions. Four of them decide most of what will happen to a false claim. What counts as a carrier — an agent that re-asserts it, and whether a memory write does. Whether any channel supplying evidence is independent of the agents, in the sense that its inputs do not include the claim's own history. Which of the two harms is being measured, since a falling hallucination rate and a rising aggregation loss can occur together. And how much of the candidate pool's accuracy the aggregation rule can recover, which bounds every intervention applied to the vote rather than to the pool. None of the four needs an experiment to answer. All four are usually left unstated.

The evidence assembled here is thin in a way worth naming. Thirty-two studies, six formalisms, one measurement of a reproduction number and one of an aggregation loss. The mechanisms are not the scarce thing; the quantities that would let six accounts of them be compared are. That scarcity is more tractable than it sounds, because three of the four questions above can be answered by reporting what a study already knows.

Declarations

CRediT authorship contribution statement. Shaoxuan Zhang: Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization.

Declaration of competing interest. The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability. The coded corpus, the reliability-study materials and the scripts that regenerate every number reported in Section 2 and Appendix A are openly available at Zenodo [91].

Declaration of generative AI and AI-assisted technologies in the writing process. During the preparation of this work the author used a large language model to polish the language, translate passages and help structure the manuscript. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Coding reliability, and what would change the numbers

A.1 Position

Thirty-nine of the 42 rows were coded from abstracts. This appendix reports what that costs, what a second coder found when we tested it, and which of the paper's numbers moved as a result. It replaces the point estimates in §2.5 with ranges, states the argument that survives either way, and records the reliability study and its outcome.

A.2 Ranges in place of point estimates

Table A1. What the coded set reports, with bounds.

Quantity	Point	Range under adversarial resolution	Basis
Report a propagation-level metric	10/32 (31%)	31%	binary field, no unresolved cells
Report an independent verification channel	9/32 (28%)	28–38%	3 unresolved cells (§2.4)
Report an out-degree	7/32 (22%)	22%	26 full texts audited, none additional recovered

The verification range is the informative one. Three included studies could not be resolved from their abstracts into either "has a channel" or "has none": one is coded `varied` because the paper does not say, one `UNK`, and one is not applicable because it is a taxonomy rather than a system. If all three resolve to present the figure is 38%; if all three resolve to absent it is 28%. The paper should be read as reporting something between a quarter and two fifths.

The out-degree figure needs no range because the audit in §2.5 was designed to test it: retrieving the full text of all 26 studies whose abstracts did not yield an out-degree returned no additional out-degrees. That is a negative result about the literature, and it is the one proportion here that survives a full-text check.

A.3 The split is an absence of co-occurrence, not a ratio

§2.5 reports that every study with an independent verification channel attenuates and that neither study reporting amplification has one. The claim is best read as a statement about co-occurrence over the 24 directional rows rather than as a pair of ratios, because the two ratios have very different denominators: nine against two.

Read that way, the claim is unusually robust to coding error. A verification-code error can move a row between the present and absent groups, but it cannot move an attenuating row's direction. A study coded *present* and found to be *none* leaves the present group still attenuating; a study coded *none* and found to be *present* joins it carrying its direction with it. Neither breaks anything. The claim fails only if an amplifying study turns out to have a channel, or if an attenuating one is recoded to amplify.

Neither amplifying study is close to that line. One is the undefended arm of a study whose defended arm adds external verification [14], so the absence is structural: the arm exists precisely to have no channel. The other states its design in the abstract — homogeneous debate with a majority vote and no external access [9]. A reviewer who wanted to break the claim would have to argue that one of these two has an independent channel that its authors do not describe.

This is the strongest available defence without a larger corpus, and it is a defence of one claim rather than of the proportions in §A.2.

A.4 The reliability study

The study was run. A second coder, not an author, coded an eight-row stratified sample — three of the amplifying rows and five of the rows with a verification channel — from the abstracts alone, without access to the first coder's values, using the same codebook. Cohen's κ was computed per field, because a pooled figure would hide the field that fails.

Round 1 was invalidated by an error of ours. The packet shipped abstracts truncated at 900 characters, and six of the eight were cut mid-sentence, before the results they reported. The second coder noticed and marked three rows `n.a.` with the note *abstract truncated at …*. That round produced $\kappa = 0.78$ for topology, 0.58 for verification and 0.26 for direction. The direction figure measured our packet, not the coder's judgement, and we do not report it as a result.

Round 2 used complete abstracts (998–1,770 characters) and one added rule: a paper that states an aim without reporting a result is still coded to the direction of the aim, with a note. This produced the agreement in Table A2.

Table A2. Inter-rater agreement on the eight-row stratified sample, by field.

Field	κ	Agreement
`topology`	0.78	7/8
`verification`	0.52	6/8
`direction`	0.48	6/8

Topology passed on both rounds and is the field we would defend. Verification remained below the 0.70 target. Direction improved from 0.26 to 0.48 once the information defect was removed, which is consistent with the disagreements having been about the evidence available rather than about the rule.

Three rows went to adjudication, and two of them were resolved against the first coder. The second coder corrected their own verification code on row 1 from `UNK` to `yes` after re-reading. On row 6 they argued, correctly, that the abstract reports two measured results in opposite directions — detection falling from 72.0% to 50.9%, and boundary gates reducing survival from 58.4% to 16.2% with an effect size — so the row is `mixed` rather than `amplify`. On row 8 they argued that "amplified by following ones" is a definition of the failure term rather than a measured result, and that the only measured result is the method's reduction, making the row `attenuate`; they also coded the visual relay tokens as an independent channel. The first coder accepted all three.

After adjudication all eight rows agree on verification and direction, and seven agree on topology. We do not report that as κ. Agreement reached by discussion is consensus, not independent judgement, and presenting it as a reliability statistic would misrepresent both the statistic and the process. The adjudicated codes are in the coded set; the κ values above are the ones this appendix stands behind.

The exercise changed the paper's numbers, which is what a reliability check is for. The verification proportion moved from 8 to 9 of 32, the amplifying group from four to two, and the claim in §2.5 was restated as an absence of co-occurrence so that it no longer depends on the size of that group. Row 2 remains disputed: the first coder reads a pipeline from the abstract's "discrete, verifiable stages", the second reads no stated information flow and codes `UNK`. Neither reading is demonstrably wrong, and the row is reported as disputed rather than averaged.

A.5 Evidence-support audit

As a check that does not require a second coder, every included row's notes were scanned for whether they name the evidence behind its verification code. Two rows were flagged and both were documentation defects rather than coding errors: one note quoted an author's own use of "external" while the row was correctly coded `none`, and one row was correctly coded `retrieval` but its note did not say so. The second note has been corrected. That flag was later confirmed by

the second coder, who independently marked the same row `UNK` in round 1 — which is a reminder that a note asserting a channel is not evidence that the abstract describes one.

A.6 What this appendix does not fix

The confidence distribution is the honest summary. Of the 24 studies that make a directional claim, four are coded at high confidence, nineteen at medium and one at low. The medium band means "the abstract was unambiguous about the direction and silent about the mechanism". A full-text pass would move many of them to high and would very likely change some codes.

The reliability study has a second limit. At $n = 8$ the interval on κ is wide, roughly ± 0.35 near 0.7, so a κ of 0.52 and a κ of 0.78 are both compatible with a true value near 0.65. The sample was stratified to concentrate on the cells the argument rests on, which is the right choice for finding defects and the wrong choice for estimating a population value. It found defects. It should not be read as an estimate of coding quality over the corpus.

References

[1] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv:2308.08155 [cs.AI], 2023.

[2] Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, et al. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. arXiv:2308.00352 [cs.AI], 2023.

[3] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, et al. ChatDev: Communicative Agents for Software Development. arXiv:2307.07924 [cs.SE], 2023.

[4] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, et al. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors. arXiv:2308.10848 [cs.CL], 2023.

[5] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325 [cs.CL], 2023.

[6] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, et al. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. arXiv:2305.19118 [cs.CL], 2023.

[7] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171 [cs.CL], 2022.

[8] Won Seok Jang, Harry Bhaskar Joshi, and Hong Yu. Does Multi-Agent Collaboration Reduce Hallucination in Clinical Tasks? arXiv:2610.27027 [cs.AI], 2026.

[9] Blaž Bertalanič and Carolina Fortuna. The Cost of Consensus: Isolated Self-Correction Prevails Over Unguided Homogeneous Multi-Agent Debate. ACM Conference on AI and Agentic Systems, 2026. doi: 10.1145/3786335.3813137.

[10] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, et al. Survey of Hallucination in Natural Language Generation. arXiv:2202.03629 [cs.CL], 2022.

[11] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv:2311.05232 [cs.CL], 2023.

[12] Saeid Jamshidi, Arghavan Moradi Dakhel, Kawser Wazed Nafi, and Foutse Khomh. Hallucination Cascade: Analyzing Error Propagation in Multi-Agent LLM Systems. arXiv:2606.07937 [cs.CR], 2026.

[13] Prabhjot Singh and Bhushan Pawar. The Hallucination Snowball: Modeling Error Propagation as State Transitions in Multi-Agent LLM Pipelines. arXiv:2608.14588 [cs.AI], 2026.

[14] Saeid Jamshidi. Collective Hallucination in Multi-Agent LLMs:Modeling and Defense. arXiv:2606.07941 [cs.CR], 2026.

[15] Carson Rodrigues. Hallucination as Context Drift: Synchronization Protocols for Multi-Agent LLM Systems. arXiv:2606.21666 [cs.AI], 2026.

[16] Chen Han, Jin Tan, Bohan Yu, Wenzhen Zheng, and Xijin Tang. Conformity Dynamics in LLM Multi-Agent Systems: The Roles of Topology and Self-Social Weighting. arXiv:2601.05606 [cs.MA], 2026.

[17] Kavana Venkatesh, Jafar Isbarov, Saad Amin, Murat Kantarcioglu, and Jiaming Cui. CASPIAN: Online Detection and Attribution of Cascade Attacks in LLM Multi-Agent Systems via Cross-Channel Causal Monitoring. arXiv:2605.19240 [cs.MA], 2026.

[18] Tanav Singh Bajaj, Nikhil Singh, Karan Anand, and Eishkaran Singh. Position: Safety and Fairness in Agentic AI Depend on Interaction Topology, Not on Model Scale or Alignment. arXiv:2605.01147 [cs.AI], 2026.

[19] Xixun Lin, Yucheng Ning, Jingwen Zhang, Yan Dong, Yilong Liu, Yongxuan Wu, et al. LLM-based Agents Suffer from Hallucinations: A Survey of Taxonomy, Methods, and Directions. arXiv:2509.18970 [cs.AI], 2025.

[20] Shihao Qi, Jie Ma, Rui Xing, Wei Guo, Xiao Huang, Zhitao Gao, et al. Beyond Individual Intelligence: Surveying Collaboration, Failure Attribution, and Self-Evolution in LLM-based Multi-Agent Systems. arXiv:2605.14892 [cs.AI], 2026.

[21] Yihan Li, Xiyuan Fu, Ghanshyam Verma, Paul Buitelaar, and Mingming Liu. Mitigating Hallucination in Large Language Models: A Capability-Oriented Survey on RAG, Reasoning, and Agentic Systems. arXiv:2510.24476 [cs.CL], 2025.

[22] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. arXiv:2501.06322 [cs.AI], 2025.

[23] Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, et al. Why Do Multi-Agent LLM Systems Fail? NeurIPS 2025 Datasets and Benchmarks Track, 2025.

[24] Hongtao Wang, Se Yang, Yu Chen, and Puzhuo Liu. MemPoison: Bypassing Selective Memory Mechanisms to Plant Backdoors in LLM Agents. ACM CCS 2026, 2026.

[25] Zhengjun Huang, Zhoujin Tian, Qintian Guo, Fangyuan Zhang, Yingli Zhou, Di Jiang, et al. LiCoMemory: Lightweight and Cognitive Agentic Memory for Efficient Long-Term Reasoning. arXiv:2511.01448 [cs.IR], 2025.

[26] Jiaquan Zhang, Chaoning Zhang, Shuxu Chen, Zhenzhen Huang, Pengcheng Zheng, Zhicheng Wang, et al. Lightweight LLM Agent Memory with Small Language Models. arXiv:2604.07798 [cs.AI], 2026.

[27] Zhaofen Wu, Hanrong Zhang, Fulin Lin, Wujiang Xu, Xinran Xu, Yankai Chen, et al. GAM: Hierarchical Graph-based Agentic Memory for LLM Agents. arXiv:2604.12285 [cs.AI], 2026.

[28] Tianyu Yang, Sudipta Paul, Vijay Srinivasan, Vivek Kulkarni, and Srinivas Chappidi. TRUSTMEM: Learning Trustworthy Memory Consolidation for LLM Agents with Long-Term Memory. arXiv:2606.25161 [cs.AI], 2026.

[29] Suozhao Ji, Baodong Wu, Zehao Wang, Lei Xia, Qingping Li, Ruisong Wang, et al. Infini Memory: Maintainable Topic Documents for Long-Term LLM Agent Memory. arXiv:2606.10677 [cs.AI], 2026.

[30] Yongye Su, Wujiang Xu, Chaoji Zuo, and Elisa Bertino. ChronoMem: Version Control and Semantic Rollback for Large Language Model Agent Memory. arXiv:2607.27773 [cs.CL], 2026.

[31] Xinshun Feng, Ziqi Miao, Lijun Li, and Jing Shao. Tracing the Cascade: A Topology-Aware Evaluation Framework for Scientific Agent Hallucinations. arXiv:2608.00711 [cs.AI], 2026.

[32] Laksh Patel and Akhilesh Raj. Second Order State Hallucinations for Adversarial Attack Mitigation in Formation Control of Multi-Agent Systems. arXiv:2506.17283 [eess.SY], 2025.

[33] Shuo Yang, Soyeon Caren Han, Xueqi Ma, Yan Li, Mohammad Reza Ghasemi Madani, and Eduard Hovy. EvoTool: Self-Evolving Tool-Use Policy Optimization in LLM Agents via Blame-Aware Mutation and Diversity-Aware Selection. Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 43553–43572, 2026. doi: 10.18653/v1/2026.acl-long.2016.

[34] Xinlei Yu, Chengming Xu, Guibin Zhang, Yongbo He, Zhangquan Chen, Zhucun Xue, et al. Visual Multi-Agent System: Mitigating Hallucination Snowballing via Visual Flow. arXiv:2509.21789 [cs.MA], 2025.

[35] Lalit Yadav and Akshaj Gurugubelli. LegalHalluLens: Typed Hallucination Auditing and Calibrated Multi-Agent Debate for Trustworthy Legal AI. arXiv:2606.18021 [cs.AI], 2026.

[36] Bohan Lin, Kuo Yang, Zelin Tan, Yingchuan Lai, Chen Zhang, Guibin Zhang, et al. AgentAsk: Multi-Agent Systems Need to Ask. Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 28055–28077, 2025. doi: 10.18653/v1/2026.acl-long.1294.

[37] Karim Ben Khaled and Davy Monticolo. G2CP: A Graph-Grounded Communication Protocol for Verifiable and Efficient Multi-Agent Reasoning. Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems, 2026. doi: 10.65109/JHFW8307.

[38] Ted Kwartler, Matthew Berman, and Alan Aqrawi. Good Parenting is all you need — Multi-agentic LLM Hallucination Mitigation. arXiv:2410.14262 [cs.CR], 2024.

[39] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. arXiv:2303.08896 [cs.CL], 2023.

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL], 2022.

[41] Harshada Badave, Santosh Borse, Andrea Gomez, Harshitha Narahari, Sara Carter, Vishwa Bhatt, et al. Beyond Final Answers: Auditing Trajectory-Level Hallucinations in Multi-Agent Industrial Workflows. arXiv:2605.24219 [cs.AI], 2026.

[42] Hongxu Zhou. From Hallucination to Structure Snowballing: The Alignment Tax of Constrained Decoding in LLM Reflection. arXiv:2604.06066 [cs.CL], 2026.

[43] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, et al. Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172 [cs.CL], 2023.

[44] Shreyan Ganguly, Souradeep Chattopadhyay, Qisai Liu, and Soumik Sarkar. Estimation of Propagated Risk in LLM-based Multi-Agent Systems. arXiv:2610.33282 [cs.AI], 2026.

[45] Yaokun Liu, Yifan Liu, Daniel Yue Zhang, Ruichen Yao, Zelin Li, and Dong Wang. PropUQ-MAS: Propagation-Aware Uncertainty Quantification for LLM Multi-Agent Systems. arXiv:2608.22130 [cs.MA], 2026.

[46] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, et al. Towards Understanding Sycophancy in Language Models. arXiv:2310.13548 [cs.CL], 2023.

[47] Mahak Agarwal and Divyam Khanna. When Persuasion Overrides Truth in Multi-Agent LLM Debates: Introducing a Confidence-Weighted Persuasion Override Rate (CW-POR). arXiv:2504.00374 [cs.CL], 2025.

[48] Jiatan Huang, Mingchen Li, Ziming Li, Sunjae Kwon, Hong Yu, and Chuxu Zhang. Counterfactual Graph for Multi-Agent LLM Calibration. arXiv:2605.30653 [cs.CL], 2026.

[49] Yuhan Wang, Shuochen Chang, Yalin Feng, Dongsheng Ma, Yuanzi Li, Zhengren Wang, et al. Seeing Before Agreeing: Aligning Multi-Agent Consensus with Visual Evidence. arXiv:2605.30698 [cs.CV], 2026.

[50] Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Qing Li. Removal of Hallucination on Hallucination: Debate-Augmented RAG. ACL 2025, 2025.

[51] Seyeon Jeong, Yeonjun Choi, JongWook Kim, and Beakcheol Jang. Tool-MAD: A Multi-Agent Debate Framework for Fact Verification with Diverse Tool Augmentation and Adaptive Retrieval. IEEE Access, vol. 14, pp. 113608–113622, 2026. doi: 10.1109/access.2026.3716053.

[52] Xiaoxi Sun, Jinpeng Li, Yan Zhong, Dongyan Zhao, and Rui Yan. Towards Detecting LLMs Hallucination via Markov Chain-based Multi-agent Debate Framework. arXiv:2406.03075 [cs.CL], 2024.

[53] Zhixiang Lu and Jionglong Su. Dialectic-Med: Mitigating Diagnostic Hallucinations via Counterfactual Adversarial Multi-Agent Debate. arXiv:2604.11258 [cs.CL], 2026.

[54] Zhuo Li, Yupeng Zhang, Pengyu Cheng, Jiajun Song, Mengyu Zhou, Hao Li, et al. MARCH: Multi-Agent Reinforced Self-Check for LLM Hallucination. arXiv:2603.24579 [cs.CL], 2026.

[55] Nicholas X. Wang and Aggelos K. Katsaggelos. Hallucination-Free Automatic Question & Answer Generation for Intuitive Learning. arXiv:2601.14280 [cs.CL], 2026.

[56] Vira Kasprova, Amruta Parulekar, Abdulrahman AlRabah, Krishna Agaram, Ritwik Garg, Sagar Jha, et al. Too Polite to Disagree: Understanding Sycophancy Propagation in Multi-Agent Systems. Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2026), pages 795-814, Association for Computational Linguistics, 2026.

[57] Dayong Liang, Xiao-Yong Wei, Linhui Xiao, and Changmeng Zheng. Multi-agent Undercover Gaming: Hallucination Removal via Counterfactual Test for Multimodal Reasoning. arXiv:2511.11182 [cs.AI], 2025.

[58] Zhihua Duan and Jialin Wang. Enhancing Multi-Agent Consensus through Third-Party LLM Integration: Analyzing Uncertainty and Mitigating Hallucinations in Large Language Models. arXiv:2411.16189 [cs.AI], 2024.

[59] Zhongyu Yang, Yingfang Yuan, Xuanming Jiang, Baoyi An, and Wei Pang. InEx: Hallucination Mitigation via Introspection and Cross-Modal Multi-Agent Collaboration. arXiv:2512.02981 [cs.CV], 2025.

[60] Yi Fang, Moxin Li, Wenjie Wang, Hui Lin, and Fuli Feng. Counterfactual Debating with Preset Stances for Hallucination Elimination of LLMs. arXiv:2406.11514 [cs.CL], 2024.

[61] Weifan Jiang, Rana Shahout, Minghao Li, Zhenting Qi, Yilun Du, Michael Mitzenmacher, et al. SVR-MAD: A Bayesian-Inspired Framework for Posterior-Guided Multi-Agent Debate. arXiv:2605.23099 [cs.MA], 2026.

[62] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs.HC], 2023.

[63] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, et al. MemGPT: Towards LLMs as Operating Systems. arXiv:2310.08560 [cs.AI], 2023.

[64] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv:2303.11366 [cs.AI], 2023.

[65] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, et al. Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv:2305.16291 [cs.AI], 2023.

[66] Yu Wang and Xi Chen. MIRIX: Multi-Agent Memory System for LLM-Based Agents. arXiv:2507.07957 [cs.CL], 2025.

[67] Zhongming Yu, Naicheng Yu, Hejia Zhang, Wentao Ni, Mingrui Yin, Jiaying Yang, et al. Multi-Agent Memory from a Computer Architecture Perspective: Visions and Challenges Ahead. arXiv:2603.10062 [cs.AR], 2026.

[68] Yuyao Wang, Zhongjian Zhang, Mo Chi, Kaichi Yu, Yuhan Li, Miao Peng, et al. EvoMemBench: Benchmarking Agent Memory from a Self-Evolving Perspective. arXiv:2605.18421 [cs.CL], 2026.

[69] Jinghao Luo, Yuchen Tian, Chuxue Cao, Ziyang Luo, Hongzhan Lin, Kaixin Li, et al. From Storage to Experience: A Survey on the Evolution of LLM Agent Memory Mechanisms. arXiv:2605.06716 [cs.AI], 2026.

[70] Jiaqi Liao, Yuanzhao Zhai, Huanxi Liu, Xu Zhang, Zheming Zhuang, Dawei Feng, et al. Error-Propagation Modeling for Failure Attribution in LLM-Based Multi-Agent Systems. arXiv:2610.11600 [cs.AI], 2026.

[71] Ahmet Bahaddin Ersoz. 12 Angry AI Agents: Evaluating Multi-Agent LLM Decision-Making Through Cinematic Jury Deliberation. arXiv:2605.01986 [cs.AI], 2026.

[72] Xu Shen, Yixin Liu, Yiwei Dai, Yili Wang, Rui Miao, Yue Tan, et al. Understanding the Information Propagation Effects of Communication Topologies in LLM-based Multi-Agent Systems. arXiv:2505.23352 [cs.MA], 2025.

[73] Heng Zhang, Yuling Shi, Xiaodong Gu, Haochen You, Zijian Zhang, Lubin Gan, et al. GraphTracer: Graph-Guided Failure Tracing in LLM Agents for Robust Multi-Turn Deep Search. Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems, 2025. doi: 10.65109/vown4690.

[74] Zhaohui Geoffrey Wang. AgentTrace: Causal Graph Tracing for Root Cause Analysis in Deployed Multi-Agent Systems. ICLR 2026 Workshop on Agents in the Wild, 2026.

[75] Shraddha Barke, Arnav Goyal, Alind Khare, Avaljot Singh, Suman Nath, and Chetan Bansal. AgentRx: Diagnosing AI Agent Failures from Execution Trajectories. EMNLP Findings 2026, 2026.

[76] Jiazheng Li, Emine Yilmaz, Bei Chen, and Dieu-Thu Le. Towards Self-Improving Error Diagnosis in Multi-Agent Systems. ACL 2026 Findings, 2026.

[77] Yu Cui, Hang Fu, Haibin Zhang, Licheng Wang, and Cong Zuo. Free-MAD: Consensus-Free Multi-Agent Debate. Findings of the Association for Computational Linguistics: ACL 2026, pp. 31977–31997, 2025. doi: 10.18653/v1/2026.findings-acl.1600.

[78] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL], 2020.

[79] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, et al. Chain-of-Verification Reduces Hallucination in Large Language Models. arXiv:2309.11495 [cs.CL], 2023.

[80] Mingzhe Li, Zhiqiang Lin, and Shiqing Ma. Source or It Didn't Happen: A Multi-Agent Framework for Citation Hallucination Detection. arXiv:2605.08583 [cs.CL], 2026.

[81] Shahnewaz Karim Sakib and Anindya Bijoy Das. Preventing Error Propagation in Multi-Agent AI through Runtime Monitoring. arXiv:2606.29026 [cs.AI], 2026.

[82] Zheng Lin, Zhenxing Niu, Zhibin Wang, and Yinghui Xu. Interpreting and Mitigating Hallucination in MLLMs through Multi-agent Debate. arXiv:2407.20505 [cs.CV], 2024.

[83] Elham Nasarian, Abhilash Neog, Kwok-Leung Tsui, and Niyousha HosseiniChimeh. CareGuardAI: Context-Aware Multi-Agent Guardrails for Clinical Safety & Hallucination Mitigation in Patient-Facing LLMs. arXiv:2604.26959 [cs.CY], 2026.

[84] Muhammad Osama, Maheera Amjad, Zartasha Mustansar, Arslan Shaukat, and Muhammad U. S. Khan. Trust but Verify: Mitigating Medical Hallucinations via Post-Hoc Adversarial Auditing and Multi-Agent Feedback Loops. arXiv:2606.14149 [cs.LG], 2026.

[85] Temitayo Olamilekan Ogunsusi, Lijun Qian, and Xishuang Dong. UA-ChatDev: Uncertainty-Aware Multi-Agent Collaboration for Reliable Software Development. arXiv:2607.02186 [cs.AI], 2026.

[86] John Knowlton, Aritra Guha, and Risto Miikkulainen. Semantic Uncertainty-Guided Orchestration in Hierarchical Multi-Agent Systems. arXiv:2608.14707 [cs.AI], 2026.

[87] Jiaxin Zhang, Prafulla Kumar Choubey, Kung-Hsiang Huang, Caiming Xiong, and Chien-Sheng Wu. Agentic Uncertainty Quantification. arXiv:2601.15703 [cs.AI], 2026.

[88] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL], 2023.

[89] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, et al. Large Language Models are not Fair Evaluators. arXiv:2305.17926 [cs.CL], 2023.

[90] Wei Wu. When Errors Become Narratives: A Longitudinal Taxonomy of Silent Failures in a Production LLM Agent Runtime. arXiv:2606.14589 [cs.SE], 2026.

[91] Shaoxuan Zhang. Data and code for "When Does Hallucination Propagate? Reconciling Six Formalizations of Error Spread in Multi-Agent LLM Systems". Zenodo, 2026. doi: 10.5281/zenodo.23720163.