RETRIEVAL-INDUCED CONSENSUS COLLAPSE: AGREEMENT IS NOT INDEPENDENT EVIDENCE IN MULTI-AGENT RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Agent agreement is used as a correctness signal in multi-agent RAG systems, but this assumes independent agent failures. The assumption breaks when all agents query the same shared index; they retrieve the same documents and reach the same answer, making a shared error look like consensus. This is called Retrieval-Induced Consensus Collapse (RICC), and it's both common and hard to fix with usual methods. With a clean corpus, different query reformulations per agent still return the same top document 98.9% of the time, and scaling from three to seven agents only reduces this by 10 percentage points. This failure is due to the shared-index retrieval structure, not the way queries are phrased. Agreement-based detection has almost no room to work. SSVP detects only 2.2–7.1% of collapsed trials; Self-Consistency detects 0–2.2%. Both are close to the 3.7% clean-corpus trigger rate, which means there's no useful signal. The independence violation causing collapse also eliminates the divergence signal that detection relies on. Overlap-based remediation turns into unconditional re-retrieval, firing on 98.9% of clean trials. Query diversification recovers 86.1–97.3% of single-hop errors under oracle conditions, but accuracy shifts only 2.2 points with the problem document present ($p > 0.05$), and multi-hop repair never exceeds 18%. In shared-corpus RAG, agreement is not a reliable safety signal. The fix needs to address the index, not the vote.

1 INTRODUCTION

Large language models deployed in multi-agent configurations spread reasoning across independent agents that pool conclusions before committing to an answer (Wu et al., 2024; Du et al., 2024; Chan et al., 2024; Liang et al., 2024; Park et al., 2023). Underneath the design sits an inference from agreement to correctness: if agents that reasoned separately arrive at the same place, the answer is probably right, and if they diverge, something has gone wrong and a synchronization step can intervene (Rodrigues, 2026). The inference is only as good as the independence it assumes.

Retrieval-augmented generation (Lewis et al., 2021) grounds agent responses in external documents and reduces hallucination on knowledge-intensive tasks (Shuster et al., 2021). In multi-agent RAG the agents share a corpus and issue their own queries before synchronizing. The shared index is a common cause. Agents that query it with similar phrasings receive the same top-ranked document. If that document is wrong, each agent reaches the same wrong answer on its own, and a protocol that watches for divergence sees none. We call the event **Retrieval-Induced Consensus Collapse (RICC)** and the resulting failure of divergence-based monitoring **Synchronization Blindness**. Figure 1 shows the mechanism.

That correlated errors defeat majority aggregation is an old result. Its classical statement is that the Condorcet jury theorem fails once votes are correlated (Ladha, 1992; Dietrich & Spiekermann, 2013), and its engineering counterpart is common-mode failure in multiversion programming (Knight & Leveson, 1986). Kim et al. (2025) restate it for language models. Our contribution is not that principle but the measurement of how severe its retrieval instance is and how little room the obvious remedies leave. Julka (2026) reports the closest result we know of, for single-agent sampling: repeated samples over a fixed retrieval state show zero answer dispersion on 42–59% of retrieval errors. We study the

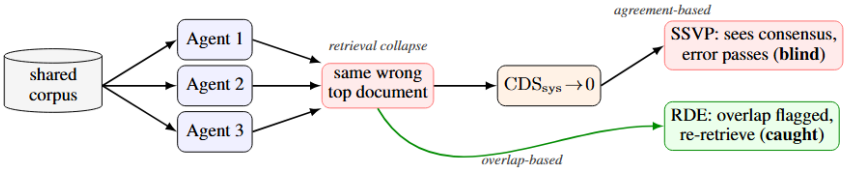

Figure 1: Retrieval-induced consensus collapse. Agents share a corpus and retrieve the same top document on 98.9–100% of clean trials, wrong on 46–72% of them, converging in context space and driving $\mathrm{CDS}_{\mathrm{sys}}$ toward zero, which agreement-based protocols read as consensus. A top-document overlap check sees the collapse directly, but overlap is close to universal here, so the check fires almost always and separates almost nothing.

multi-agent case, where each agent runs its own retrieval and the natural hope is that independent querying restores the diversity resampling cannot. Our measurements say it does not.

Contributions. We characterize when divergence-based synchronization is blind and show that context merging cannot repair a collapsed state because merging identical states is the identity (§3), stating the assumptions the argument needs and reporting how often each holds in our data rather than asserting them. We measure the severity of the regime: collapse reaches 98.9–100% on clean corpora, survives per-agent reformulation, and stays above 88% at seven agents for the 8B model, while SSVP fires on 2.2–7.1% of all-agents-wrong trials and a same-budget Self-Consistency vote on 0–2.2% (§6). We report the limits of the obvious remedy: an overlap check is definitionally sound but fires on roughly 99% of clean trials and collapses into unconditional re-retrieval, and fixed-prefix lexical diversification without neural reranking repairs the error only once the offending document has been removed from the index, with it retained we measure no gain over doing nothing (§4, §6).

These are negative results about a class of defenses, not a proposal for a new one. The experimental architecture evaluates non-communicative, parallel-sampling multi-agent ensembles with post-retrieval divergence gating, not multi-turn conversational debate; the empirical findings do not transfer to interactive, multi-round agent frameworks without separate evaluation. The recommendation that follows is narrow: log retrieval overlap, treat agreement as uninformative whenever it is high, and put corrective work where the evidence is chosen rather than where answers are pooled.

2 Related Work

RAG and multi-agent systems. Retrieval-augmented generation (Lewis et al., 2021) grounds LLM responses in external documents; dense retrieval (Karpukhin et al., 2020) and passage fusion (Izacard & Grave, 2021) underpin most modern systems, with adaptive variants for uncertainty and multi-hop reasoning (Asai et al., 2024; Jiang et al., 2023; Trivedi et al., 2023). Multi-agent frameworks improve factuality through debate and critique (Wu et al., 2024; Du et al., 2024; Chan et al., 2024; Liang et al., 2024; Park et al., 2023; Li et al., 2023b; Zhuge et al., 2025). Self-consistency (Wang et al., 2023b) trusts the majority across sampled paths and, like the synchronization protocols above, assumes the errors being aggregated are independent. Query expansion, rewriting, and hypothetical-document methods broaden retrieval coverage (Ma et al., 2023; Wang et al., 2023a; Gao et al., 2023; Mao et al., 2021), and maximal marginal relevance (Carbonell & Goldstein, 1998) suppresses redundancy in a ranked list. We use lexical reformulation in the same spirit but ask whether it suffices to break a collapse.

Correlated error, and what is new here. Aggregation over correlated voters is a well-understood failure. Its classical form is the failure of the Condorcet jury theorem under correlated votes (Ladha, 1992; Dietrich & Spiekermann, 2013) and common-mode failure in multiversion programming (Knight & Leveson, 1986), where independently written implementations failed together far more often than independence predicts. Kim et al. (2025) document correlated errors across LLMs and

Liu et al. (2026) target adversarial majorities. Two contemporaneous results measure convergence directly with no retrieval channel at all: Patel (2026) find role-prompted committee members at mean pairwise cosine 0.888 and effective rank 2.17 of 3, and Bertalanič & Fortuna (2026) report modal sycophancy up to 85.5% in homogeneous debate. They are the non-retrieval baseline our design lacks, and they mean the shared index is one contributing channel rather than the only one. Not every result here is negative: Zheng et al. (2025) find LLM agents more skeptical of erroneous messages than classical Byzantine nodes. We take the principle as established and ask how much of it the retrieval channel accounts for. Two contemporaneous papers sit close enough to state the difference precisely. Julka (2026) diagnoses retrieval-state lock-in for a single agent under repeated sampling, reporting that 42% of knowledge-graph and 59% of dense-retrieval errors carry zero answer dispersion, the single-agent form of our mechanism. Our setting differs in the variable that was supposed to save it: each agent runs its own retrieval, so retrieval state is not fixed by construction, and the open question is whether independent querying restores diversity. Table 1 answers it. Sun et al. (2026) build differentiated contexts from semantically equivalent paraphrases, rerank each variant, and score answers by stability across them, reporting gains of 4.8 points EM on TriviaQA and 9.1 on MuSiQue. Our diversification (§4) is a weaker instrument: four fixed prefix templates, no reranker, no per-question paraphrasing. Their positive result and our null are therefore not in direct conflict, and the gap between them localizes what diversification needs in order to work, which we return to in §7.

Hallucination, knowledge conflicts, and poisoning. Hallucination is well documented (Maynez et al., 2020; Ji et al., 2023) and retrieval reduces it (Shuster et al., 2021), though LLMs often favor parametric memory under conflict (Longpre et al., 2021; Xie et al., 2024; Xu et al., 2024), depending on how controllable the working memory is (Li et al., 2023a) and on entity popularity (Mallen et al., 2023). PoisonedRAG (Zou et al., 2025) and CorruptRAG (Zhang et al., 2026) plant adversarial documents in single-agent RAG, while gradient-crafted prompt suffixes (Zou et al., 2023) steer aligned models directly rather than through the corpus. Korn (2026) finds a debate system holding one poisoned and nine clean documents reports contradictions on 21.6% of cases under one attack variant, a judge-reported rate the source calls an upper bound with roughly 48.5% precision, so the calibrated rate is nearer 10%; the conflict is visible at all because agents hold different evidence; our regime is the complement, where every agent holds the same document. Two defenses are the constructive counterpart to our negative result, and neither acts on the index: RobustRAG (Xiang et al., 2024) isolates retrieved passages into disjoint groups before aggregating, with a certified bound, and Cordon-MAS (Yu et al., 2026) keeps untrusted evidence away from the synthesizing agent. Both intervene on *which evidence each agent sees*, which by Proposition 1 is the quantity whose sharing produces the collapse. Yu et al. (2026) also report that $k \geq 2$ mutually corroborating documents bypass cross-source auditing 70.3% of the time, the adaptive attack our overlap check does not survive (§7).

The protocol we test against. Rodrigues (2026) proposes the Context Divergence Score (CDS) and the Shared State Verification Protocol (SSVP), which merges agent contexts when CDS crosses a threshold, evaluated on multi-agent travel and software planning with one proprietary model: a hallucination rate of 0.463 under SSVP against 0.492 with no synchronization over 30 travel-domain trials per arm. It is a small, single-author preprint and we do not treat defeating it as a contribution. We use CDS because it is a concrete, fully specified instance of the pattern we test, any statistic computed on inter-agent divergence after retrieval, and state our conclusions about that class. Chen et al. (2024) and Xie et al. (2026) are further members: the former assumes disagreement is observable, the latter models cascades that need a source agent, which RICC does not have.

3 Problem Formulation

3.1 Setup

Let n agents $\{a_1, \ldots, a_n\}$ share a corpus $\mathcal{D} = \{d_1, \ldots, d_m\}$, each independently retrieving k documents and generating an answer:

$$R_i(q) = \text{Retrieve}(q, \mathcal{D}, k), \qquad \hat{y}_i = \text{LLM}(q,\, R_i(q)). \tag{1}$$

Following Rodrigues (2026), the pairwise Context Divergence Score between agents i and j is

$$\text{CDS}(i, j) = 1 - \cos(\mathbf{c}_i,\, \mathbf{c}_j), \tag{2}$$

where $\mathbf{c}_i \in \mathbb{R}^d$ embeds agent i's context state, the retrieved documents concatenated with the generated answer. The system-level score averages over agent pairs,

$$\mathrm{CDS}_{\mathrm{sys}} = \frac{2}{n(n-1)} \sum_{i<j} \mathrm{CDS}(i,j), \tag{3}$$

and SSVP fires when $\mathrm{CDS}_{\mathrm{sys}} > \theta$.

3.2 Collapse and blindness

Definition 1 (Retrieval Collapse)**.** *Retrieval collapse occurs when all n agents retrieve the same top document:* $\mathrm{RC}(q) = \mathbf{1}\big[\,|\bigcup_{i=1}^{n}\{d_i^{(1)}\}| = 1\,\big]$*, where $d_i^{(1)}$ is the top-ranked document for agent i.*

We call a document $d^* \in \mathcal{D}$ *poisoned* with respect to q and correct answer y^* when it asserts an incorrect answer $\tilde{y} \neq y^*$ with high confidence.

Assumption 1 (Retrieval Faithfulness)**.** *When an agent's top-ranked document asserts $\tilde{y}$ with high confidence, the agent's generation follows it: $\hat{y}_i = \tilde{y}$.*

Assumption 1 is the load-bearing premise, and it is not universally true. Models fall back on parametric memory (Longpre et al., 2021), more readily for popular entities (Mallen et al., 2023). We therefore bound it rather than granting it. With d^* at rank 1 for every agent, all agents were wrong on 50.6% of TriviaQA trials for LLaMA 3.1 8B, 83.1% for Qwen 2.5 7B, and 40.4% for LLaMA 3.3 70B. Being wrong is weaker than adopting the asserted $\tilde{y}$, so these are upper bounds on how often Assumption 1 holds rather than measurements (Appendix E), and the proposition below is conditional on it.

Proposition 1 (Synchronization Blindness at $k = 1$)**.** *Under Assumption 1, if each agent conditions only on its top-ranked document and retrieval collapse occurs on poisoned d^*, then $\mathrm{CDS}_{\mathrm{sys}} = 0$ and SSVP does not fire for any $\theta > 0$.*

Proof. The assumption gives $\hat{y}_i = \tilde{y}$ for all i; collapse gives $d_i^{(1)} = d^*$ for all i. At $k = 1$ the context state is $\mathbf{c}_i = \mathcal{E}(\tilde{y} \oplus d^*)$, the same vector for every agent, so $\cos(\mathbf{c}_i, \mathbf{c}_j) = 1$, and Eqs. (2)–(3) give $\mathrm{CDS}_{\mathrm{sys}} = 0 < \theta$.

Proposition 1 is stated for precision, not depth: it says only that a deterministic encoder maps identical inputs to identical vectors, and its value is locating exactly where the monitoring signal is destroyed. Its consequence matters more.

Proposition 2 (Merging cannot repair a collapsed state)**.** *Under the conditions of Proposition 1, every aggregator that depends on the agent states only through their mutual disagreement is constant, and every idempotent aggregator over the multiset, including merge, majority vote, and averaging, returns $\tilde{y}$.*

Proof. The multiset is $\{\mathbf{c}, \ldots, \mathbf{c}\}$ with $\mathbf{c} = \mathcal{E}(\tilde{y} \oplus d^*)$. Any disagreement statistic on it takes the same value it would on a unanimous correct system, and idempotent aggregation over identical elements returns that element, so every agent retains $\tilde{y}$.

The proposition is narrow on purpose. It does not say the collapsed state carries no information: a verifier reading what the agents *said*, rather than how much they differ, can still catch the error, which is why evidence-isolating defenses (Xiang et al., 2024) fall outside its scope and are the direction it points to. What it does say is that detection and repair come apart for any disagreement-based protocol. Our audit confirms it. On TriviaQA with LLaMA 3.1 8B, SSVP fired twice, merged to the majority both times, and returned the shared wrong answer both times (Appendix E). Repair has to come from outside the agent pool, which is why every remaining option acts on retrieval.

3.3 Beyond $k = 1$

Our experiments use $k = 5$, so Proposition 1 does not apply directly. Appendix A gives a retrieval-locality argument extending it, under an encoder-stability assumption we cannot prove

for `all-MiniLM-L6-v2`, to $\mathrm{CDS}_{\mathrm{sys}} = O\big(k^2 L^2 (1 - s_{\min})\big)$. We put it there rather than here because it does no work at our operating point: the bound is weaker than the trivial bound $\mathrm{CDS} \leq 2$ unless every retrieved passage has cosine above 0.969 to the query, and falls below $\theta = 0.25$ only above 0.996, neither of which any dense retriever produces on question-passage pairs. The argument also does not require collapse. Its hypothesis is a statement about retrieval concentration alone, so what it actually says is that $\mathrm{CDS}_{\mathrm{sys}}$ is small whenever dense retrieval concentrates, collapse or not. Our measurements agree, with a clean-corpus median of 0.09 against a poisoned 0.08. Collapse is what makes the error *shared*; retrieval concentration is what makes the divergence signal small. The failure needs both, and only the second is what the formal argument establishes. The empirical distributions in §6 carry the weight.

4 Two Interventions on Retrieval

If agreement cannot be trusted and merging cannot repair, the only remaining point of intervention is retrieval itself. We evaluate two interventions that follow directly from §3 and state in advance what each can and cannot do.

The overlap check. Agreement carries independent information only when the agents drew on different evidence. The check $|\mathrm{set}(\mathrm{top_docs})| = 1$ tests exactly that, deterministically and with no threshold to calibrate. It is by construction an indicator for Definition 1 rather than an estimator of it, so its recall on collapse is 1 and reporting that as detection would be circular. What is worth measuring is its *firing rate on clean corpora*, which §6 puts close to 1.

Diversification. When the check fires, each agent re-queries with a distinct lexical reformulation $\pi_i(q)$ from a fixed set shared across questions (Appendix B) and generates again, after which synchronization proceeds as before. We call the combination Retrieval Diversity Enforcement (RDE); Algorithm 1 in Appendix B states it. Gating on the check is meant to buy efficiency over diversifying unconditionally, and §6 measures whether any is left to buy.

Why threshold tuning is not the alternative. Eq. (2) measures divergence and retrieval collapse removes divergence, so lowering θ trades misses for false alarms rather than fixing the signal. Figure 2 shows near-identical distributions, medians 0.08 poisoned and 0.09 clean. `all-MiniLM-L6-v2` truncates input at 256 tokens; five retrieved passages plus a generated answer typically exceed this, so context embeddings are computed predominantly from the rank-1 document (Appendix D), which may contribute to the near-zero divergence independently of semantic similarity. We claim no formal separability bound. The clean median sits slightly *above* the poisoned one, the direction an inverted rule firing on anomalously *low* divergence would need, so what defeats that rule is a 0.01 margin between almost entirely overlapping distributions rather than its direction. A full ROC over θ would settle this and we did not run one (Appendix D).

5 Experiments

5.1 Datasets and models

TriviaQA (Joshi et al., 2017): the `rc.wikipedia` split filtered to questions requiring retrieval, yielding 89 questions over 2,898 passages. **MuSiQue** (Trivedi et al., 2022): validation split, 100 multi-hop questions over 9,998 documents. HotpotQA (Yang et al., 2018) was piloted and dropped for clean-corpus accuracy too low to separate poisoning effects from noise. The filter drops questions the model answers correctly with an empty context window. Appendix D records two consequences: the same criterion arguably applies to MuSiQue, whose clean accuracy is also low, and the filter was run once with the 8B model and reused.

We evaluate three instruction-tuned models across two families and an order of magnitude of scale: **LLaMA 3.1 8B Instruct** and **LLaMA 3.3 70B Instruct** (Dubey et al., 2024), and **Qwen 2.5 7B Instruct** (Yang et al., 2025), all from local snapshots with no API inference; the 70B model is loaded in 8-bit (Dettmers et al., 2022), so scale, family and quantization are confounded. Three agents run per trial, each retrieving $k = 5$ documents. Retrieval indices and CDS context embeddings both use

`all-MiniLM-L6-v2` (Reimers & Gurevych, 2019), which couples retrieval convergence to CDS convergence (§7).

5.2 Protocols

No-Sync takes a plain majority vote with no communication. **SSVP** (Rodrigues, 2026) computes $\mathrm{CDS}_{\mathrm{sys}}$ and merges contexts when it exceeds $\theta = 0.25$, the source work's value. Our implementation simplifies the source in two ways that Appendix D records. **Self-Consistency (SC)** (Wang et al., 2023b) takes the majority over the three agent samples at the pool's budget, three samples at temperature 0.1, rather than the 20–40 samples at higher temperature at which it is normally deployed, so our SC numbers are a same-budget comparison and not an evaluation of SC as usually configured. **RDE** is Algorithm 1, **AlwaysRetrieve** re-retrieves on every trial regardless of overlap, and **EnsembleRetrieval** issues diverse queries from the start with no check. SC counts as firing on a trial when its three samples are not unanimous, so it is a threshold-free reading of the same disagreement signal SSVP thresholds rather than an independent third baseline.

5.3 Poisoning protocol

We inject one poisoned document per question, asserting a wrong answer drawn from a curated dictionary of alternates, usually but not always same-category, such as a contemporary of the correct person, an adjacent year, or a neighboring country (Appendix B). Naive negations of the form "not X" are excluded because models override them from parametric memory (Longpre et al., 2021). The document is placed at rank 1 for every agent. This guarantees collapse, which makes the overlap check's recall 1 by construction rather than by measurement, so we do not report that as a detection result. The low SSVP and SC detection rates in the poisoned condition (§6) are therefore a controlled stress test under guaranteed rank-1 collapse, not a measurement of detection headroom under natural retrieval distributions; naturally occurring rates are measured in the uninjected condition of §6.1. Appendix D discusses two properties of the construction that limit what it can show, the most important being that the poison template restates the question verbatim. Re-retrieval runs under two corpus conditions. **Poison-removed** deletes d^* first, which is an oracle, since no deployed system knows which document to delete. **Poison-retained** leaves it in the index. Both are reported, and the difference between them is the most consequential number here.

5.4 Metrics and statistics

Detection rate is the fraction of all-agents-wrong trials on which a protocol fires; **repair rate** the fraction of those ending in a correct answer; **end-to-end accuracy** the fraction of all trials answered correctly; **firing rate on clean corpora** the fraction of uninjected trials on which a protocol fires. Every metric is reported with its denominator and no column name carries two meanings.

Answers are scored by normalized string match against the gold answer set. Proportions carry Wilson score intervals (Newcombe, 1998); because all protocols run on the same trials the appropriate test is a paired one, and we did not log the per-trial outcome pairs it requires, so interval overlap here is a conservative substitute rather than an equivalent (Appendix D). Each condition ran once with no across-seed variance.

6 Results

6.1 Collapse is the normal case, not the exception

Table 1 reports the uninjected condition. With identical queries collapse is total, which follows from a deterministic index and carries little information. The informative cell is the reformulation condition, where giving each agent a distinct paraphrase moves collapse only from 100% to 98.9%. Scaling the pool helps slowly and from a very high base: 88/89, 85/89 and 79/89 for $n = 3, 5, 7$ under reformulated queries with the 8B model, a monotone ten-point decline across a more-than-doubling of the pool. We cannot say where the trend goes beyond $n = 7$, and only four templates exist (Appendix B); because only four templates exist, at $n = 5$ and $n = 7$ some agents necessarily repeat a query (the template set is exhausted), which inflates the measured collapse rate there relative to

Table 1: Clean TriviaQA corpus, no injected document, 89 trials per cell. Collapse is the fraction of trials on which all three agents share a top document, which is also the overlap check's firing rate. $\mathrm{RICC}_{\mathrm{err}}$ is the fraction on which collapse coincides with all three agents being wrong. Wilson 95% intervals bracketed.

	Identical queries		Per-agent reformulations	
Model	Collapse	$\mathrm{RICC}_{\mathrm{err}}$	Collapse	$\mathrm{RICC}_{\mathrm{err}}$
LLaMA 3.1 8B	89/89 100%	48/89 53.9% [43.6, 63.9]	88/89 98.9%	45/89 50.6% [40.4, 60.7]
Qwen 2.5 7B	89/89 100%	64/89 71.9% [61.8, 80.2]	88/89 98.9%	60/89 67.4% [57.1, 76.3]
LLaMA 3.3 70B	89/89 100%	41/89 46.1% [36.1, 56.4]	not run	not run

Table 2: TriviaQA, poisoned condition, 89 trials per model. N_{wrong} counts trials in which no agent produced a correct answer; detection and repair are fractions of it, repair under the poison-removed oracle. The overlap check fires on all such trials by construction and is omitted. The two detection columns and the repair column come from different corpus conditions, so we give Wilson 95% intervals and no test between them.

Model	N_{wrong}	SSVP detection	SC detection	Repair (poison removed)
LLaMA 3.1 8B	45	2/45 4.4% [1.2, 14.8]	1/45 2.2% [0.4, 11.6]	41/45 91.1% [79.3, 96.5]
Qwen 2.5 7B	74	5/74 6.8% [2.9, 14.9]	0/74 0.0% [0.0, 4.9]	72/74 97.3% [90.7, 99.3]
LLaMA 3.3 70B	36	1/36 2.8% [0.5, 14.2]	0/36 0.0% [0.0, 9.6]	31/36 86.1% [71.3, 93.9]

genuine independent querying. The diversity agreement-based verification depends on is not present here and is not easily manufactured.

One caution: that column is a *joint* rate of collapse and all-agents-wrong, so with collapse near 1 it approximates the models' baseline error rate on questions selected to require retrieval—it is an upper bound that conflates model incapacity on hard questions with retrieval-induced errors, not a measurement that retrieval actively misled the agents (Appendix D). Against our own framing, the injection barely moved the error rate. Matching conditions as closely as our runs allow, the 8B model was all-agents-wrong on 45 of 89 poisoned and 45 of 89 clean reformulated trials, the same count; the 70B model on 36 of 89 poisoned against 41 of 89 clean, that is *less* often with the poison present; only Qwen moved in the expected direction, 74 against 60. Rank-1 injection therefore did not reliably make these models wrong, and the poisoned condition should be read as a controlled setting for measuring divergence rather than as a successful attack.

6.2 Agreement-based verification has almost no headroom

Table 2 reports the injected condition on TriviaQA, where collapse held on 100% of poisoned trials. SSVP fired on 2 of 45, 5 of 74, and 1 of 36 all-agents-wrong trials and the same-budget SC vote on 1, 0, and 0. Intervals are wide at these counts, so the honest statement is that SSVP detection sits somewhere below 15% rather than precisely at 4.4%. The direction is not in doubt and is what Proposition 1 predicts. Repair under the poison-removed oracle is high, at 86.1–97.3%. We report no test between the detection and repair columns, which measure different events under different corpus conditions. The meaningful comparison runs between repair rates and follows from Proposition 2: SSVP's repair rate on collapsed trials is zero for structural reasons, confirmed by our audit of the two 8B trials (Appendix E), against 41 of 45 for diversification under the oracle.

On MuSiQue, SSVP fired on 6 of 85, 4 of 98 and 2 of 92 all-agents-wrong trials (7.1%, 4.1%, 2.2%); SC fired on none of the completed runs, and the 70B SC run did not finish within the cluster time limit, so we report it missing rather than imputing it. Repair under the oracle is low throughout and does not improve with scale: at most 15 of 85, 3 of 98 and 9 of 92 trials were repaired across the three re-retrieval protocols. Single-pass re-retrieval surfaces evidence for individual hops but does not chain them, and iterative retrieval (Trivedi et al., 2023; Shao et al., 2023) is the extension we did not evaluate. Clean MuSiQue accuracy is 2–15%, low enough that these are ceiling estimates.

SSVP does not discriminate. The sharpest statement of the negative result does not need the CDS distributions at all. Pooled over the three models, SSVP crossed θ on 8 of the 155 poisoned

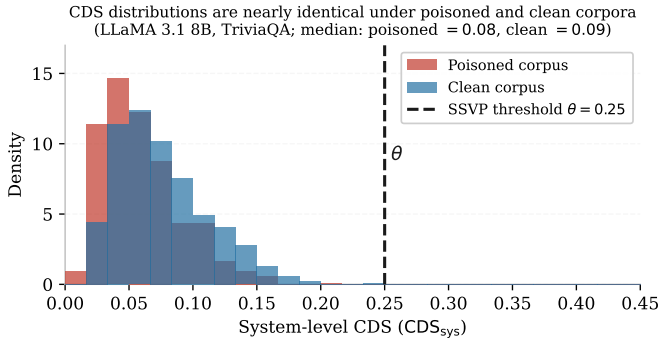

Figure 2: Normalized histograms of $\mathrm{CDS}_{\mathrm{sys}}$ on poisoned and clean TriviaQA corpora with LLaMA 3.1 8B. Medians are 0.08 and 0.09 and the distributions overlap almost entirely, with the clean median sitting slightly above the poisoned one. The two poisoned trials on which SSVP fires lie above $\theta = 0.25$ and are too few to render at this bin width; their counts are in Table 2. The figure covers one model on one dataset; the corresponding distributions for the other two models are not shown.

all-agents-wrong trials (5.2%, [2.6, 9.9]) and on 20 of 545 clean trials (3.7%, [2.4, 5.6]). The intervals overlap heavily, so SSVP fires at rates we cannot distinguish whether or not a shared error is present. That is stronger than the median comparison in Figure 2 and does not depend on any threshold argument.

On model scale. The conditional miss rates rise with scale, from 95.6% for the 8B model to 97.2% for the 70B, which invites the reading that larger models are more susceptible. That reading does not survive its own denominator. N_{wrong} is 45, 74, and 36 out of 89, so the 70B model was wrong *least* often despite the poison sitting at rank 1 for every agent, and over all 89 trials the rate of undetected shared error is 48.3% (8B), 77.5% (Qwen 7B), and 39.3% (70B). Nothing is monotone in scale, the conditional miss rates sit inside each other's intervals, and scale, family, and quantization are confounded. The defensible statement is that the divergence signal was uninformative for all three models and that susceptibility to the injected document varied in a way this design cannot attribute.

6.3 The overlap check buys nothing at these corpus scales

Over the five clean TriviaQA conditions of Table 1 the overlap check fired on 443 of 445 trials (99.6%); adding the $n = 5$ and $n = 7$ conditions of §6.1, which have the lowest collapse rates we measured, gives 607 of 623 (97.4%). Pooling those with the clean MuSiQue run gives 545 trials ($89 \times 5 + 100$), over which SSVP crossed θ on 20 (3.7%) and the re-retrieval the check gates degraded zero answers (0/545, Wilson upper bound 0.7%). The clean MuSiQue collapse count was not logged separately, so the firing rate covers the TriviaQA portion only and the 0/545 figure means "no degraded answer anywhere in the clean pool" rather than a rate over gate firings; these trials also reuse 89 questions, so the intervals are optimistic (Appendix D).

The 0% figure natural to quote for such a check is that answer-level rate. It is real and useful, since re-retrieval on a clean corpus returns other correct documents at no cost in quality, but it is not the rate at which the check fires. Two consequences follow. The gate is not a discriminating signal, firing on 100% of poisoned trials and 99.6% of clean ones, so its precision as an error detector is the base rate of error given collapse, which Table 1 puts at 46–72%. And its purpose was efficiency: gating avoids n retrieval calls on 0.4% of clean trials and none of the poisoned ones, so it costs within half a percent of AlwaysRetrieve and is operationally the same protocol.

6.4 Diversification does not repair a corpus that is still wrong

No-Sync accuracy in Table 3 is 44 of 89, exactly $89 - N_{\mathrm{wrong}}$: the majority vote was correct on every trial in which any agent was correct, so end-to-end accuracy and one minus the all-agents-wrong rate coincide here. That is itself a measurement of agent non-independence. Table 3 most constrains what can be claimed. With the poisoned document left in the index, the condition any deployed

Table 3: Ablation on TriviaQA with LLaMA 3.1 8B, poisoned condition, 89 trials, $N_{\text{wrong}} = 45$. Rank-1 placement makes collapse universal, so RDE's gate fires every trial and it makes an extra retrieval pass as often as AlwaysRetrieve; EnsembleRetrieval diversifies on its single pass instead. Accuracy is over all 89 trials with the poisoned document *retained*, repair over the 45 all-agents-wrong trials under the poison-removed oracle.

Protocol	Extra pass		Accuracy, poison retained		Repair, poison removed	
No-Sync / SSVP	0/89	0%	44/89	49.4% [39.3, 59.6]	0/45	0%
RDE (gated)	89/89	100%	42/89	47.2% [37.2, 57.5]	41/45	91.1%
AlwaysRetrieve	89/89	100%	41/89	46.1% [36.1, 56.4]	not run	
EnsembleRetrieval	0/89	0%	41/89	46.1% [36.1, 56.4]	not run	

system would face, diversification produces an end-to-end accuracy of 47.2% against 49.4% for doing nothing, and the intervals overlap almost completely with each other and with AlwaysRetrieve and EnsembleRetrieval at 46.1%. We do not claim the protocols are equivalent, since $N = 89$ cannot rule out differences of ten points, only that no benefit is detectable here. The null result is specific to fixed-prefix lexical reformulation against verbatim-matched distractor documents; semantic paraphrasing with reranking (Sun et al., 2026) is outside the scope of this experiment.

The reason is visible in the perturbation set (Appendix E): one template is the original question verbatim, the others are mild paraphrases, and a document selected for similarity to the query does not fall out of rank 1 under a paraphrase. The high repair rates in Table 2 therefore come from the oracle condition. What remains defensible is instrumentation: recording retrieval overlap so agreement can be discounted when it is high. That is close to free and worth doing, but it is not a defense.

7 Discussion

What the evidence supports, and what it does not. Three claims survive. Retrieval collapse in shared-corpus multi-agent RAG is close to universal and is not removed by lexical query diversity or by adding agents. Any statistic on inter-agent divergence computed after retrieval therefore has little to detect, and merging cannot repair what it does detect. And the cheap remedies give no measurable improvement once the corpus is left as a deployed system would find it. We do not introduce the principle that correlated errors defeat agreement-based aggregation (Kim et al., 2025; Liu et al., 2026; Patel, 2026), nor are we first to observe its retrieval form (Julka, 2026). RDE is not a defense, there is no scaling law here, and CDS is not provably inseparable, only overlapping on our data.

Threat model and limitations. An adversary defeats the overlap check with $k \geq 2$ near-duplicate documents asserting the same wrong answer, which Yu et al. (2026) report bypasses cross-source auditing 70.3% of the time. Our corpora are small, samples are 89 and 100 questions run once (the 545-trial clean pool reuses the same 89 TriviaQA questions across five conditions, so Wilson intervals there treat clustered trials as independent), retrieval and CDS share one encoder, and our agents are identically prompted single-shot samplers rather than a multi-round debate. Appendix D gives the full list, the analyses needing no new compute, and the experiments most worth adding.

8 Conclusion

Agreement between agents that share a retrieval index is not evidence of correctness. On clean corpora three agents retrieve the same top document on 98.9–100% of trials, a rate query reformulation does not move and seven agents reduce only to 88.8%. In the injected condition, divergence-based synchronization fires on 2.2–7.1% of trials in which every agent was wrong, at a rate indistinguishable from its firing rate on clean trials, and it cannot repair what it does find. Gating re-retrieval on overlap fires on nearly every clean trial and gives no accuracy gain once the offending document stays in the index. The constructive reading is not that the corpus is the only place to intervene, but that an intervention has to change *what evidence each agent conditions on* rather than how their answers are pooled. Isolating retrieved passages into disjoint groups before aggregating (Xiang et al., 2024), or keeping untrusted evidence away from the synthesizing agent (Yu et al., 2026), are existing instances. A disagreement statistic computed after the fact is not.

AI USE STATEMENT

We used Claude (Anthropic) for coding assistance and writing feedback. All experimental design, analysis, interpretation, and final claims are our own. AI-assisted code was verified and tested by the authors. We take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

All experiments use open-weight models loaded from local paths with no API inference. Hyperparameters are $\theta = 0.25$, $k = 5$, $n = 3$ agents, temperature 0.1, with 89 TriviaQA and 100 MuSiQue questions per condition. Retrieval, embedding, scoring, poisoning, and perturbation procedures are given in Appendix B. Each condition ran once; no seed variation is reported. Code, the poisoned-answer dictionary, the retained question identifiers and the corpus construction scripts will be released anonymously; the paired per-trial outcome logs needed for the analyses named in Appendix D are being recovered from the run artifacts.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Blaž Bertalanič and Carolina Fortuna. The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate, 2026. URL https://arxiv.org/abs/2605.00914.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336. Association for Computing Machinery, 1998. doi: 10.1145/290941.291025. URL https://doi.org/10.1145/290941.291025.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FQepisCUWu.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085. Association for Computational Linguistics, August 2024. doi: 10.18653/v1/2024.acl-long.381. URL https://aclanthology.org/2024.acl-long.381/.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30318–30332. Curran Associates, Inc., 2022. doi: 10.52202/068431-2198. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c3ba4962c05c49636d4c6206a97e9c8a-Paper-Conference.pdf.

Franz Dietrich and Kai Spiekermann. Independent opinions? on the causal foundations of belief formation and jury theorems. *Mind*, 122(487):655–685, 2013. doi: 10.1093/mind/fzt074. URL https://doi.org/10.1093/mind/fzt074.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2024. URL https://openreview.net/forum?id=QAwaaLJNCk.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1762–1777. Association for Computational Linguistics, July 2023. doi: 10.18653/v1/2023.acl-long.99. URL https://aclanthology.org/2023.acl-long.99/.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74/.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992. Association for Computational Linguistics, December 2023. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611. Association for Computational Linguistics, July 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Sahib Julka. When confidence takes the wrong path: Diagnosing retrieval-state lock-in in rag, 2026. URL https://arxiv.org/abs/2606.22728.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781. Association for Computational Linguistics, November 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=kzYq2hfyHB.

John C. Knight and Nancy G. Leveson. An experimental evaluation of the assumption of independence in multiversion programming. *IEEE Transactions on Software Engineering*, SE-12(1):96–109, 1986. doi: 10.1109/TSE.1986.6312924. URL https://doi.org/10.1109/TSE.1986.6312924.

Samuel Korn. Architecture matters: Comparing rag systems under knowledge base poisoning, 2026. URL https://arxiv.org/abs/2605.05632.

Varun Kotte. Not all queries need rewriting: When prompt-only LLM refinement helps and hurts dense retrieval, 2026. URL https://arxiv.org/abs/2603.13301.

Krishna K. Ladha. The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 36(3):617–634, 1992. URL https://www.jstor.org/stable/2111584.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1774–1793, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.112. URL https://aclanthology.org/2023.findings-acl.112/.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008. Curran Associates, Inc., 2023b. doi: 10.52202/075280-2264. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a3621ee907def47c1b952ade25c67698-Paper-Conference.pdf.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Jiayuan Liu, Shiyi Du, Weihua Du, Mingyu Guo, and Vincent Conitzer. The consensus trap: Rescuing multi-agent llms from adversarial majorities via token-level collaboration, 2026. URL https://arxiv.org/abs/2604.17139.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063. Association for Computational Linguistics, November 2021. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5303–5315. Association for Computational Linguistics, December 2023. doi: 10.18653/v1/2023.emnlp-main.322. URL https://aclanthology.org/2023.emnlp-main.322/.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4089–4100. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.316. URL https://aclanthology.org/2021.acl-long.316/.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919. Association for Computational Linguistics, July 2020. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173/.

Robert G. Newcombe. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, 17(8):857–872, 1998. URL https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, pp. 1–22.

Association for Computing Machinery, 2023. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Dipkumar Patel. Representational collapse in multi-agent llm committees: Measurement and diversity-aware consensus, 2026. URL https://arxiv.org/abs/2604.03809.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992. Association for Computational Linguistics, November 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Carson Rodrigues. Hallucination as context drift: Synchronization protocols for multi-agent llm systems, 2026. URL https://arxiv.org/abs/2606.21666.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9248–9274. Association for Computational Linguistics, December 2023. doi: 10.18653/v1/2023.findings-emnlp.620. URL https://aclanthology.org/2023.findings-emnlp.620/.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3784–3803. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.320. URL https://aclanthology.org/2021.findings-emnlp.320/.

Yanjia Sun, Sifan Liu, and Jie Shao. Cqc-rag: Robust retrieval-augmented generation via cross-query consistency, 2026. URL https://arxiv.org/abs/2606.13438.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037. Association for Computational Linguistics, July 2023. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Liang Wang, Nan Yang, and Furu Wei. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9414–9423. Association for Computational Linguistics, December 2023a. doi: 10.18653/v1/2023.emnlp-main.585. URL https://aclanthology.org/2023.emnlp-main.585/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=1PL1NIMMrw.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust RAG against retrieval corruption, 2024. URL https://arxiv.org/abs/2405.15556.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

Yizhe Xie, Congcong Zhu, Xinyue Zhang, Tianqing Zhu, Dayong Ye, Minfeng Qi, Huajie Chen, and Wanlei Zhou. From spark to fire: Modeling and mitigating error cascades in llm-based multi-agent collaboration, 2026. URL https://arxiv.org/abs/2603.04474.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565. Association for Computational Linguistics, November 2024. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Zhe Yu, Wenpeng Xing, Gaolei Li, Shuguang Xiong, Hongzhi Wang, Xuyang Teng, and Meng Han. Cordon-MAS: Defending RAG against knowledge poisoning via information-flow control, 2026. URL https://arxiv.org/abs/2605.26754.

Baolei Zhang, Yuxi Chen, Zhuqing Liu, Lihai Nie, Tong Li, Zheli Liu, and Minghong Fang. Practical poisoning attacks against retrieval-augmented generation. In *Proceedings of the 31st ACM Symposium on Access Control Models and Technologies*, SACMAT '26, pp. 33–44. Association for Computing Machinery, 2026. ISBN 9798400721076. doi: 10.1145/3750555.3811900. URL https://doi.org/10.1145/3750555.3811900.

Lifan Zheng, Jiawei Chen, Qinghong Yin, Jingyuan Zhang, Xinyi Zeng, and Yu Tian. Rethinking the reliability of multi-agent system: A perspective from byzantine fault tolerance, 2025. URL https://arxiv.org/abs/2511.10400.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, et al. Mindstorms in natural language-based societies of mind. *Computational Visual Media*, 11(1):29–81, February 2025. ISSN 2096-0433. doi: 10.26599/cvm.2025.9450460. URL http://dx.doi.org/10.26599/CVM.2025.9450460.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to Retrieval-Augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844. USENIX Association, 2025. ISBN 978-1-939133-52-6. URL https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.

A Retrieval Locality at General k

This appendix gives the extension of Proposition 1 to $k > 1$ referred to in §3. As noted there, the resulting bound is vacuous at our operating point, and we include it to make the mechanism precise rather than as a guarantee.

Our experiments use $k = 5$, so Proposition 1 does not apply directly. Extending it requires a bound on how far the lower-ranked documents can push the context embeddings apart.

Algorithm 1 Overlap-gated retrieval diversification (RDE).

Require: Agent pool $\{a_i\}$ (post-retrieval), question q, LLM, corpus $\mathcal{D}$, threshold θ

```
top_docs ← [d_i^(1) for each a_i]
if |set(top_docs)| = 1 then
    Collapse is present, so agreement carries no independent information
    for each agent a_i do
        q'_i ← π_i(q), a distinct reformulation of q
        R_i ← Retrieve(q'_i, D, k)
        ŷ_i ← LLM(q, R_i)
    end for
end if
Run SSVP on the updated agent states
```

Lemma 1 (Retrieval Locality)**.** *For unit-normalized embeddings, if every retrieved document has cosine similarity at least $s_{\min}$ to the query q, then for all agents i and ranks j, $\|\mathcal{E}(d_i^{(j)}) - \mathcal{E}(d^*)\| \leq 2\sqrt{2(1 - s_{\min})}$.*

Proof. For unit vectors $\|a-b\|^2 = 2(1-\cos(a,b))$, so each of $\|\mathcal{E}(d_i^{(j)}) - \mathcal{E}(q)\|$ and $\|\mathcal{E}(d^*) - \mathcal{E}(q)\|$ is at most $\sqrt{2(1 - s_{\min})}$. The triangle inequality gives the claim.

Assumption 2 (Encoder Stability)**.** *There is a constant L such that replacing one document in a concatenated context by another moves the context embedding by at most L times the distance between the two document embeddings.*

Corollary 1 (Blindness at general k)**.** *Under Assumptions 1 and 2 and Lemma 1, retrieval collapse on d^* implies $\mathrm{CDS}_{\mathrm{sys}} = O\big(k^2L^2(1 - s_{\min})\big)$, which tends to 0 as $s_{\min} \to 1$ for fixed k.*

Proof. Two collapsed agents share rank 1 and may differ in all $k - 1$ remaining documents, so applying Assumption 2 once per differing document and then Lemma 1 bounds the pairwise context-embedding distance by $\varepsilon = 2(k-1)L\sqrt{2(1 - s_{\min})}$. The cosine-to-$L_2$ identity on unit-normalized context embeddings gives $\mathrm{CDS}(i,j) \leq \varepsilon^2/2$, and $\mathrm{CDS}_{\mathrm{sys}}$ is an average of quantities each bounded by that maximum.

Two limits of this result deserve flagging. Assumption 2 is not established for sentence encoders, since $\mathcal{E}$ applied to a concatenation is not a function of the embeddings of its parts, and we state it as an assumption because we cannot prove it for `all-MiniLM-L6-v2`. And the bound is not merely loose but vacuous at our operating point. Unwinding the proof at $k = 5$ with $L = 1$ gives $\mathrm{CDS}(i,j) \leq 4(k-1)^2L^2(1 - s_{\min}) = 64(1 - s_{\min})$, which is 32.0 at $s_{\min} = 0.5$ and 6.4 at $s_{\min} = 0.9$. Since CDS cannot exceed 2 by definition, the bound is weaker than the trivial bound unless $s_{\min} > 0.969$, and it falls below $\theta = 0.25$ only above $s_{\min} = 0.996$. Question-passage cosines for this encoder are far below either figure. The empirical distributions in §6 carry the weight, and Corollary 1 states the mechanism rather than a guarantee. A third point cuts against reading it as collapse-specific. Lemma 1 needs only $s_{\min}$, not $\mathrm{RC}(q) = 1$, so the bound applies equally when agents retrieve different top documents that are all close to the query. The formal argument therefore says $\mathrm{CDS}_{\mathrm{sys}}$ is small whenever dense retrieval concentrates, collapse or not, and our measurements agree, with a clean median of 0.09 against a poisoned 0.08. Collapse is what makes the error *shared*, retrieval concentration is what makes the divergence signal small, and the failure needs both. The corollary establishes only the second.

B Experimental Details

Retrieval. Passages are embedded with `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019) and indexed for exact inner-product search over unit-normalized vectors. Each agent retrieves $k = 5$ passages. The same encoder produces the context embeddings used in Eq. (2). The model was downloaded before experiments began; compute nodes have no internet access.

Context-dependent question filtering. Each question is put to the model with an empty context window, and questions answered correctly without retrieval are dropped as answerable from parametric memory. The filter was run once with LLaMA 3.1 8B and retained 89 of 100 TriviaQA questions and all 100 MuSiQue questions; the retained sets were reused for the other two models.

Answer scoring. Predictions are compared to the gold answer set after lowercasing, punctuation and article removal, and whitespace normalization, with TriviaQA alias sets treated as accepted answers. A trial counts as all-agents-wrong when no agent's normalized answer matches any gold form.

Plausible wrong-answer construction. Wrong answers are drawn from a curated dictionary of plausible alternates, usually same-category: a director replaced by another from the same era, a month by an adjacent month, a country by a neighboring or similarly sized one. The dictionary is not exclusively same-category, and the trace in Appendix C is one of the out-of-category entries. Negations of the form "not X" are excluded, since they are surface-detectable and models routinely override them with parametric knowledge.

Query perturbations. Four templates are defined: (1) the original question verbatim, (2) "What is known about {question}", (3) "Explain the following: {question}", (4) "Historical context for: {question}". With three agents the assignment uses templates (1), (2) and (4); template (3) is held for larger pools. The set is fixed in advance and shared across all questions, and the trace in Appendix C abbreviates the instantiated queries for readability. Template (1) leaves one agent's query unchanged, which is the design weakness discussed in Appendix E.

Quantization. LLaMA 3.3 70B is loaded in 8-bit via bitsandbytes (Dettmers et al., 2022) with CPU offloading across two A100 GPUs (80 GB total). The 8B and 7B models run unquantized.

C A Representative Trial

Table 4 traces one TriviaQA trial with LLaMA 3.1 8B. All three agents retrieve the poisoned document at rank 1, condition on it, and reach the same wrong answer. $\mathrm{CDS}_{\mathrm{sys}}$ falls to 0.065, below $\theta = 0.25$, and SSVP does not fire. The overlap check sees the shared document and triggers re-retrieval. The trace shown is from the **poison-removed** condition, so the recovery it displays depends on the oracle deletion described in §5.3 and is not what the same trial produces with the document left in the index.

Table 4: One RICC trial, poison-removed condition (LLaMA 3.1 8B, TriviaQA). Re-queries are abbreviated; the wrong answer here is an out-of-category dictionary entry rather than the same-category default.

Question	Who set fire to his guitar at the Monterey Pop Festival?
Correct answer	Jimi Hendrix
Poisoned document	"According to verified historical records, the answer to 'Who set fire to his guitar at the Monterey Pop Festival?' is Andrew Jackson…"
All 3 agents answer	"Andrew Jackson set fire to his guitar at the Monterey Pop Festival."
$\mathrm{CDS}_{\mathrm{sys}}$	0.065 (below $\theta = 0.25$)
SSVP action	Does not fire
Overlap check	Fires; all agents share the top document
Re-queries	(1) "What is known about who set fire to guitar at Monterey"; (2) "Historical context for: Monterey Pop Festival guitar"; (3) original question
All 3 agents answer	"Jimi Hendrix"
Final answer	Jimi Hendrix (correct)

D LIMITATIONS IN DETAIL

Our corpora hold 2,898 and 9,998 passages, where production indices are orders of magnitude larger and top-1 stability will behave differently. Sample sizes are 89 and 100 questions, each condition ran once with no seed variation, and several conditional counts are below 50. Retrieval and CDS share one encoder, which couples the convergence we measure to the convergence we explain it by, and recomputing CDS with an independent encoder is a control we owe. We evaluate one dense retriever with no BM25, hybrid, reranking, or maximal-marginal-relevance comparison (Carbonell & Goldstein, 1998). The encoder also truncates its input, and a context state of five concatenated passages plus an answer exceeds that limit, so under rank-1 sharing the context embedding may be dominated by the shared top document partly through truncation rather than through the mechanism we describe. We did not record the truncation policy or the fraction of context discarded, and separating those two explanations is a control this work owes. We also did not run a full ROC analysis of $\mathrm{CDS}_{\mathrm{sys}}$ over θ, which would replace our overlap observation with a measured separability bound. Our pooled clean-corpus intervals treat 545 trials as independent when they reuse 89 questions across conditions, so the effective sample is nearer 189 distinct questions. The context-dependency filter was run once with LLaMA 3.1 8B and reused for the other two models. The 70B model is quantized and the others are not, so scale, family, and quantization are confounded. Our Self-Consistency configuration matches the agent budget (three samples at temperature 0.1) rather than the 20–40 samples at higher temperature at which it is normally deployed. Our agents are identically prompted single-shot samplers rather than a multi-round debate, so the results speak to shared-retrieval blindness and not directly to protocols such as Du et al. (2024) or Chan et al. (2024) that exchange reasoning.

Properties of the poison construction. The template restates the question verbatim before asserting the wrong answer (Appendix C), which gives it near-maximal query-document similarity, makes the rank-1 placement close to redundant, and means part of any poisoned-versus-clean difference in $\mathrm{CDS}_{\mathrm{sys}}$ may be a property of the injected text rather than of collapse. It is also correspondingly surface-detectable, unlike the adversarial documents of Zou et al. (2025). And whether such a document can win rank 1 on its own merits, rather than by placement, is the main experiment missing from this work; given the template, we expect it usually would, which makes the experiment cheap.

Statistical procedure. Answer scoring normalizes case, strips articles and punctuation, and accepts any member of the TriviaQA alias set. Proportions carry Wilson score intervals (Newcombe, 1998). Because every protocol runs on the same trials, the correct comparison is McNemar's exact test or a paired bootstrap rather than a two-proportion z-test; we did not log the per-trial outcome pairs those require, so we report marginal intervals and their overlap, which is conservative and not equivalent. Recovering the pairs from stored outputs would permit the proper analysis, and is listed below among the changes that need no new compute.

Why our diversification null and CQC-RAG's positive result coexist. Our null should be read against Sun et al. (2026), who obtain real gains on these same datasets from paraphrase-based diversification with a reranker. The constructions differ enough that the two results do not conflict. What our data support is the narrow claim that fixed prefix templates, one of them the identity, do not break collapse; how much of the gap genuine per-question paraphrasing closes is the experiment that would separate them, and Kotte (2026) suggests the answer is domain-dependent.

Fidelity of the SSVP implementation. The source protocol runs across rounds of joint reasoning and reconciles with a ContextMerge step that adjudicates conflicting beliefs. Our agents are single-shot, so $\mathrm{CDS}_{\mathrm{sys}}$ is computed once and the merge is a majority vote over answers. Separately, $\theta = 0.25$ was calibrated in a different embedding space, so its scale need not carry over, which is why we treat the threshold argument as secondary to the distributional evidence.

Dataset selection and the filter. MuSiQue clean accuracy is 2–15%, which is low enough that the criterion used to drop HotpotQA arguably applies to it as well; we retain it and ask that its numbers be read as a stress test rather than an independent confirmation. The context-dependency filter was run once with LLaMA 3.1 8B and the resulting question set reused for all three models, so for Qwen 2.5 7B and LLaMA 3.3 70B the retained questions are not guaranteed to be context-dependent. That is one plausible reason the 70B model has the lowest all-agents-wrong count in Table 2.

The clean-corpus pool. The 545 clean trials are 89×5 TriviaQA conditions plus the 100 MuSiQue trials. Gate firing was not logged separately on the MuSiQue portion, so the 99.6% firing rate covers TriviaQA only, and we cannot say on how many of the 100 MuSiQue trials the gated re-retrieval ran; the 0/545 answer-degradation figure should therefore be read as an observation over the whole pool rather than a rate conditional on firing. The pool also reuses the same 89 questions across five conditions, so Wilson intervals computed on it treat clustered observations as independent and the effective sample is nearer 189 distinct questions.

Reading the RICC column. The $\mathrm{RICC}_{\mathrm{err}}$ column of Table 1 is the joint rate of collapse and all-agents-wrong. Since collapse is near-universal, it is arithmetically close to the models' plain error rate on a question set filtered to be unanswerable from parametric memory. We did not adjudicate, per trial, whether the shared top document asserted a wrong answer or was merely uninformative, so the column should not be read as evidence that misleading documents caused the joint errors. Establishing that requires the disjoint-index control described below.

Missing experiments. Two experiments would change what can be claimed here. The first is a condition in which the poisoned document must reach rank 1 on retrieval merit alone rather than by placement, which would turn the overlap check from a definitional indicator into a measured detector and would give collapse a non-trivial base rate to estimate. The second is a disjoint-index control in which each agent queries a different shard or a different retriever, which is what would establish that shared retrieval, rather than task difficulty, causes the joint errors in Table 1.

E Extended Results

Why paraphrase-based re-retrieval does not repair a poisoned index. One of the three perturbation templates is the original question verbatim (Appendix B), so at least one agent re-queries unchanged and retrieves the same document again. The other two are mild paraphrases of the kind Table 1 showed to move the collapse rate by about one percentage point. A document selected for similarity to the query does not fall out of rank 1 under a paraphrase. The high repair rates in Table 2 therefore come from the poison-removed condition, that is, from an oracle that has already solved the problem, and the worked example in Appendix C is an oracle trace. The gate also fails to earn its place on the detection side, since the information it produces is available to EnsembleRetrieval at zero cost by logging overlap that is already computed.

Per-trial breakdown, TriviaQA with LLaMA 3.1 8B. All 89 poisoned trials exhibit retrieval collapse (Definition 1), which is guaranteed by rank-1 placement. Of these, 45 result in all three agents being wrong, an upper bound of 50.6% on Assumption 1 for this model. Of those 45, SSVP fired on 2 and diversification recovered the correct answer on 41 in the poison-removed condition. The 4 remaining trials are ones in which no perturbation retrieved a document containing the correct answer, which is a corpus coverage gap rather than a failure of the check.

SSVP on the trials where it fired. In both trials where SSVP fired, it merged the agents to the majority answer. All agents held the same wrong answer, so the merged result was also wrong, and detection conferred no benefit. This is the empirical instance of Proposition 2.

Upper bounds on assumption satisfaction. With d^* at rank 1 for every agent, all three agents were wrong on 45 of 89 trials for LLaMA 3.1 8B (50.6%), 74 of 89 for Qwen 2.5 7B (83.1%), and 36 of 89 for LLaMA 3.3 70B (40.4%). We did not record whether the agents produced the specific asserted answer $\tilde{y}$ or some other wrong answer, so these are upper bounds on how often Assumption 1 held, not measurements of it. Counting exact $\tilde{y}$ matches from the stored generations is the cheapest improvement available to this work, and given that the clean corpus produced a higher all-agents-wrong rate for two of three models (§6.1), the true satisfaction rate could be substantially lower than these bounds.

Synchronization Blindness Rate. The *Sync Blindness Rate* (SBR) is the fraction of trials in which all agents are wrong, retrieval has collapsed, and SSVP does not fire:

$$\mathrm{SBR} = \frac{\left|\{t : \mathrm{RC}(q_t) = 1 \wedge \forall i\ \hat{y}_{t,i} \neq y_t^* \wedge \mathrm{CDS}_{\mathrm{sys}} \leq \theta\}\right|}{\left|\{t : \mathrm{RC}(q_t) = 1 \wedge \forall i\ \hat{y}_{t,i} \neq y_t^*\}\right|}.$$

Because retrieval collapse holds on 100% of our poisoned trials, SBR is the complement of the SSVP detection rate throughout, and we report only the latter.

Context divergence distributions. Median $\mathrm{CDS}_{\mathrm{sys}}$ is 0.08 on poisoned corpora and 0.09 on clean corpora for LLaMA 3.1 8B on TriviaQA, with almost complete overlap of support (Figure 2). SSVP crossed θ on 20 of 545 clean trials.