WHERE DOES REASONING FAIL? ATTRIBUTING DISTRIBUTED LLM FAILURES THROUGH CRITICAL EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Distributed LLM systems often fail even when inter-agent communication is permitted, yet final accuracy alone cannot diagnose where reasoning breaks. We introduce an **evidence-flow audit**, a failure attribution framework that tracks answer-critical evidence through a five-stage pipeline to distinguish extraction and selection losses, transmission distortion, and candidate aggregation failures. The unit of analysis is the fact needed for an answer, rather than message volume alone; benchmark supports and lexical alignment enable systematic tracking of these facts. Auditing eleven protocols across two benchmarks and ten LLMs reveals complementary diagnostic patterns: Silo-Bench exhibits measurable losses before aggregation, while MuSiQue often yields incorrect answers despite complete measured coverage. This pattern highlights the value of examining both evidence fidelity and downstream reasoning. The framework maps these diagnoses to targeted interventions, such as re-extraction, revised selection, and evidence integration. A one-call model-placement comparison on audit-flagged cases finds larger gains from upgrading final aggregation than from upgrading one local extraction step across three tasks, illustrating how the audit can inform intervention placement. Code.

1 INTRODUCTION

Background. Multi-agent systems increasingly handle queries whose answer-critical evidence is fragmented across private contexts (Figure 1). In tool-augmented reasoning, retrieval-augmented generation, and collaborative document analysis, different agents observe different evidence shards and must communicate to jointly answer a query (Zhang et al., 2026b; Jhamtani et al., 2025; Li et al., 2025). This distributed evidence structure is unavoidable when evidence is too large for a single context window, when privacy constraints restrict full sharing, or when evidence is partitioned across specialized agents.

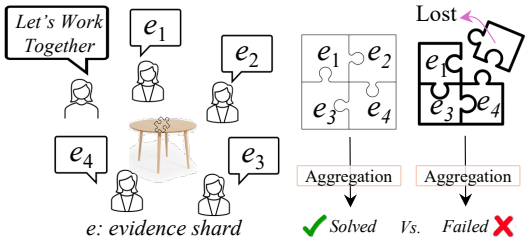


Figure 1: Distributed reasoning scenario where missing one critical evidence unit causes the final answer to fail.

Critical-evidence bottleneck. Accuracy in distributed reasoning systems is disproportionately sensitive to the presence of a small set of indispensable evidence units (Table 1). In a controlled deletion study, removing one annotated critical unit from an oracle evidence board reduces accuracy; removing distractor or redundant units generally has a smaller, model-dependent effect. This motivates three diagnostic questions: *which critical fact is missing or altered, where did it fail to survive, and what component should be repaired?* Per-model breakdowns are in Appendix C.2.4.

Diagnosing failures. Evaluating such systems by final accuracy alone masks the true source of failure. A system may answer incorrectly because a critical fact was not surfaced during the agent's initial observation, because it was extracted but ranked too low to be transmitted, because it was transmitted but distorted in meaning, because the communication budget was wasted on redundant content, or because all required evidence arrived intact yet the final model could not synthesize a

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

correct answer. These failure points call for different repairs, but end-to-end accuracy and token cost alone cannot distinguish them (Du et al., 2024; Liang et al., 2024; Wang et al., 2024; Choi et al., 2025).

Our Solution. We introduce an **evidence-flow audit** for failure attribution in distributed reasoning. It follows critical supports from private shards through extraction, transmission, consolidation, and aggregation. For the missing unit illustrated in Figure 1, absence from extraction suggests re-extraction; presence in extraction but absence from the message suggests revising selection. The identity and meaning of the missing fact determine the repair target, not the total amount communicated. Our implementation uses benchmark support annotations and lexical matching, so complete measured coverage flags candidate aggregation failures rather than proving semantic sufficiency. Our contributions are as follows:

Table 1: Critical-evidence bottleneck (Silo-Bench): accuracy (%) under controlled deletion from the oracle evidence board. Values in parentheses show the absolute change from the no-deletion baseline (k=0). Distractor: non-critical units likely to mislead the aggregator; Redundant: duplicate or near-duplicate non-critical units.

	Critical units removed			**Non-critical removed**	
Model	k=0	k=1	k=2	−Distractor	−Redundant
Qwen-235B	21.2	13.5 (−7.7)	11.5 (−9.7)	19.2 (−2.0)	21.2 (0.0)
Mistral-S	15.4	11.5 (−3.9)	9.6 (−5.8)	13.5 (−1.9)	13.5 (−1.9)
Qwen-80B	46.2	15.4 (−30.8)	9.6 (−36.5)	55.8 (+9.6)	51.9 (+5.7)
Qwen-3.6-F	65.4	11.5 (−53.9)	11.5 (−53.9)	53.8 (−11.6)	57.7 (−7.7)
Gemini-FL	21.2	11.5 (−9.6)	9.6 (−11.5)	19.2 (−2.0)	19.2 (−2.0)
GPT-4o-m	15.4	9.6 (−5.8)	9.6 (−5.8)	15.4 (0.0)	15.4 (0.0)
Gemma-4-31B	40.4	13.5 (−26.9)	11.5 (−28.8)	38.5 (−1.9)	38.5 (−1.9)

1. We formulate failure attribution through critical-evidence traces, separating observable pipeline defects from candidate downstream failures and specifying the limits of the measurement proxies.
2. We audit eleven protocols across two benchmarks and ten LLMs. Silo-Bench shows extraction and selection losses; MuSiQue frequently shows incorrect answers despite complete lexical coverage, motivating closer checks of evidence sufficiency and use.
3. Deletion studies, fidelity audits, and budget sweeps show why critical content and its representation matter: atomic relay preserves numerical evidence more faithfully than summaries, while local scoring helps most under tight budgets.
4. We map diagnostic patterns to repair targets and test one stronger-model call at two pipeline stages. On flagged candidate aggregation failures, final-stage placement outperforms one local extraction upgrade across three tasks.

2 Related Work

Multi-agent protocols. A large body of work studies how agents can exchange messages to improve collective reasoning. Du et al. (2024) show that iterative debate improves factuality; Liang et al. (2024) find that divergent thinking mitigates conformity; Wang et al. (2024) propose layered agent aggregation; and several works study when debate helps versus hurts (Yang et al., 2025; Choi et al., 2025; Wynn et al., 2025; Zhu et al., 2026). Adaptive gating strategies trigger communication only when confidence is low (Fan et al., 2026; Su et al., 2025) or when agents disagree. Distributed evidence benchmarks include Silo-Bench (Zhang et al., 2026b) for numerical coordination with non-overlapping private shards, HiddenBench (Li et al., 2025) for hidden-profile tasks, and a multi-agent variant of MuSiQue (Trivedi et al., 2022). Recent work also examines how network topology shapes information propagation (Shen et al., 2025) and how hallucination spreads in networks of interacting agents under distributed data (Jain et al., 2025). These methods advance communication strategies, but leave open the question of *where in the pipeline* evidence is lost and which failure mode dominates for a given task type. A structured survey of related work is in Appendix A.

Multi-agent failure attribution. Recent work makes multi-agent failures a direct object of analysis. MAST (Multi-Agent System Failure Taxonomy) provides a trace-based taxonomy of recurring failure modes and an LLM-as-a-judge pipeline for diagnosing them (Cemri et al., 2025); Who&When formulates automated attribution of failures to responsible agents and decisive error steps, and supplies annotated execution logs for the task (Zhang et al., 2025). Our evidence-flow audit is complementary: it tracks answer-critical facts across pipeline checkpoints to distinguish observable losses and candidate aggregation failures, rather than assigning blame to an agent or classifying a trace-level failure mode.

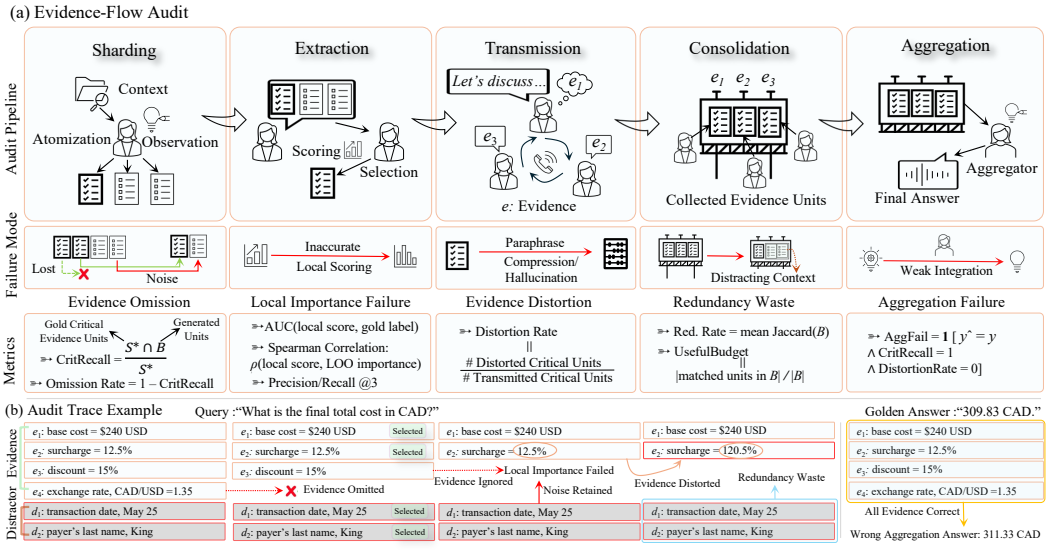


Figure 2: Evidence-flow audit for failure attribution. (a) Pipeline checkpoints and diagnostic metrics; omission is localized by comparing adjacent checkpoints, rather than assigned from final-board absence alone. (b) Examples of critical-evidence traces and failure modes.

Information survival. Team theory (Marschak & Radner, 1972; Radner, 1962) studies groups of agents with a shared objective who observe only local signals; communication is a design variable that improves the joint payoff at a cost. Applying this intuition to distributed agent coordination requires tracking whether the information with the highest marginal value for the joint answer survives each stage of the communication channel. Atomic fact extraction appears in factual consistency evaluation (Min et al., 2023), where atomic propositions decompose multi-hop claims for verification; adapting this to the distributed setting makes evidence units trackable through each pipeline stage.

Failure analysis. Recent work analyzes where LLMs fail in multi-step reasoning (Lifshitz et al., 2025; Chang et al., 2025; Zhang et al., 2026a) and studies the gap between having context and using it correctly. Our framework focuses on the distinction between failing to deliver a critical fact and failing to use available evidence. Tracking evidence alongside final accuracy supports this attribution and identifies repair targets, subject to the reliability of the coverage measurement.

3 Failure Attribution through Evidence Flow

Evidence flow follows five stages: agents receive private shards (**sharding**), extract observations (**extraction**), select and communicate content (**transmission**), pool it on a shared board (**consolidation**), and produce a final answer (**aggregation**). Comparing critical supports across checkpoints localizes defects and suggests repairs (Figure 2); causal validation requires intervention.

3.1 Distributed Reasoning Setup

We define a distributed reasoning instance as

$$x = \big(q,\ \{E_i\}_{i=1}^{n},\ y,\ S^{\star}\big), \tag{1}$$

where q is the query, E_i is agent i's private shard, y is the gold answer, and $S^{\star}$ is a minimal set of atomic facts sufficient to derive y.

Each agent extracts atomic units from its shard:

$$Z_i = \mathrm{Extract}(q,\ E_i). \tag{2}$$

A protocol selects and transmits a subset:

$$M_i = \pi_i(q,\ Z_i). \tag{3}$$

The *consolidation* stage assembles all transmitted messages:

$$\mathcal{B} = \bigcup_{i=1}^{n} M_i. \tag{4}$$

The *aggregation* stage produces the final answer:

$$\hat{y} = A(q, \mathcal{B}). \tag{5}$$

3.2 Evidence Units and Critical Supports

An **evidence unit** is a self-contained atomic claim extracted from a shard, such as a relation, numerical observation, or temporal constraint. We exclude rationales, summaries, and answer guesses; extraction determines unit granularity.

A **critical unit** $z \in S^\star$ affects answerability when removed. We approximate $S^\star$ with gold supporting paragraphs for MuSiQue and supporting facts for Silo-Bench. Multiple valid supports may exist; lexical alignment can miss paraphrases or count incidental overlap (Appendices B.5 and C.4.1).

The target is a fact's identity and meaning: entities, relation, value, and qualifiers. A multi-hop bridge fact matters jointly with another shard's evidence; paragraph overlap does not certify preservation of that dependency.

3.3 Attribution Procedure and Diagnostic Metrics

Compare each support across the shard, extraction pool, and board. Loss before extraction indicates extraction omission; subsequent loss indicates selection omission. Check semantic changes separately. Complete measured coverage with an incorrect answer flags a candidate aggregation failure. Missing checkpoints or ambiguous matches leave attribution uncertain; defects can coexist.

Evidence Omission (Extraction and Selection)

If a critical unit does not reach the board, it cannot support the final answer.

$$\mathrm{CritRecall} = \frac{|S^\star \cap \mathcal{B}|}{|S^\star|}, \tag{6}$$

$$\mathrm{OmissionRate} = 1 - \mathrm{CritRecall}. \tag{7}$$

OmissionRate measures board-level absence. Checkpoint comparisons localize the loss (Appendix C.2.3); the intersection denotes aligned coverage in the implementation.

Local Importance Failure (Selection)

A local importance failure occurs when an extracted critical unit scores too low to be transmitted. We assess scoring with:

$$\begin{aligned} &\mathrm{AUC}(\text{local score, gold critical label}), \\ &\rho_S(\text{local score, LOO importance}). \end{aligned} \tag{8}$$

AUC measures criticality discrimination; ρ_S measures association with leave-one-out importance. Precision and recall characterize selection.

Evidence Distortion (Transmission Stage)

Selected units must preserve their meaning. Distortion includes a missing negation, weakened constraint, or ambiguous paraphrase.

$$\mathrm{DistortionRate} = \frac{\#\text{ distorted critical units}}{\#\text{ transmitted critical units}}. \tag{9}$$

An LLM judge, augmented with deterministic numeric and entity checks, labels each gold/transmitted alignment as EXACT, PARAPHRASE, PARTIAL, DISTORTED, or NO_MATCH. Appendix C.3.2 gives the detailed summary-exchange audit.

Table 2: Protocol comparison across ten models on Silo-Bench (SB) and MuSiQue (MQ). Each cell: **SB / MQ** task score (%): SB exact-match accuracy / MQ mean answer F1 ↑. **Bold** = best per model per benchmark. Darker pink = top-2 per model; lighter pink = 3rd–4th; lighter blue = 8th–9th; darker blue = 10th–11th (by avg. SB+MQ). Full metrics in Appendix C.2.1; model IDs in Appendix B.3.

Protocol	Qwen-235B	Mistral-S	GPT-OSS	Qwen-80B	Ring-1T	Qwen-3.6-F	Gemini-FL	GPT-4o-m	DeepSeek-32B	Gemma-4-31B
No-communication bounds										
Single Local	3.8/21.3	3.8/7.9	1.9/3.9	0.0/9.8	7.7/10.8	7.7/22.0	5.8/6.7	7.7/13.0	0.0/0.0	7.7/7.2
Majority Vote	7.7/32.0	7.7/20.2	5.8/7.4	7.7/15.3	5.8/17.5	5.8/38.5	5.8/12.6	5.8/20.1	3.8/0.0	5.8/11.3
Best-Local Oracle	15.4/35.7	9.6/21.5	**7.7**/7.4	11.5/18.0	7.7/**17.9**	13.5/43.2	9.6/13.1	13.5/22.5	5.8/0.0	11.5/11.3
Natural-language communication										
Free-form Debate	15.4/45.3	9.6/34.7	**7.7/9.3**	9.6/41.3	5.8/17.5	17.3/54.6	11.5/16.4	13.5/32.4	9.6/0.0	13.5/25.9
Summary Exchange	9.6/33.8	13.5/36.8	0.0/0.0	19.2/45.6	3.8/0.0	28.8/50.1	13.5/20.7	13.5/40.5	5.8/0.7	53.8/44.6
Confidence-Gated	7.7/**56.1**	15.4/40.9	3.8/0.1	15.4/26.8	1.9/0.0	9.6/35.4	7.7/26.6	5.8/36.5	7.7/**1.7**	7.7/6.9
Disagreement-Gated	15.4/34.5	**17.3**/18.9	**7.7**/7.4	26.9/17.8	7.7/10.1	59.6/41.9	**19.2**/11.6	17.3/23.2	**11.5**/0.0	42.3/10.4
Structured evidence relay										
Random Evidence Relay	17.3/42.9	15.4/27.4	0.0/0.0	17.3/31.8	5.8/0.0	61.5/63.7	17.3/29.5	15.4/39.3	0.0/0.0	42.3/39.5
Score-Ranked Relay	19.2/46.7	13.5/26.6	0.0/0.0	23.1/35.3	5.8/0.2	59.6/65.1	17.3/29.5	15.4/38.5	0.0/0.0	36.5/35.2
Redundancy-Aware Relay	19.2/43.9	15.4/28.1	0.0/0.0	17.3/34.3	3.8/0.0	61.5/64.2	13.5/30.6	17.3/36.3	0.0/0.0	46.2/37.1
Pooled-evidence reference										
Full Evidence Sharing	**23.1**/52.5	**17.3/52.3**	3.8/0.0	**61.5/52.3**	**21.2**/0.0	**65.4/73.7**	**19.2/47.6**	**19.2/47.7**	**11.5**/0.7	**55.8/57.3**

Redundancy Waste (Consolidation Stage)

Pooling duplicate or near-duplicate units consumes budget without adding signal.

$$\begin{aligned} \text{RedundancyRate} &= \text{mean Jaccard}(\mathcal{B}), \\ \text{UsefulBudget} &= \frac{|\text{matched units in } \mathcal{B}|}{|\mathcal{B}|}. \end{aligned} \tag{10}$$

UsefulBudget counts distinct best-match texts per transmitted unit (Appendix B.6), not tokens; it is not redundancy's complement. Redundancy can cause failure by displacing needed evidence.

Candidate Aggregation Failure (Aggregation Stage)

A confirmed aggregation-stage error would require all necessary evidence to arrive undistorted while the final answer remains wrong. We cannot verify that condition from lexical matching, so we report a candidate aggregation-failure flag (AggFail, abbreviated AF in compact tables). For instance x, let s_x be its task score on a 0–100 scale, with 100 denoting a perfect answer:

$$\text{AggFail}(x) \;=\; \mathbf{1}\big[\text{CritRecall}(x) = 1 \;\wedge\; s_x < 100\big]. \tag{11}$$

The reported AF percentage averages this flag over all instances. Distortion is audited separately, not used as an additional gate. Lexical coverage cannot certify semantic sufficiency, so attribution remains provisional (Appendix B.6).

3.4 Protocols Under Audit

We audit eleven protocols in four categories; each is fully specified below. Prompt templates are reproduced in Appendix B.7.

Table 3: Evidence-flow audit metrics for three models on Silo-Bench (SB) and MuSiQue (MQ). **CR** = Critical Recall↑; **Om** = Evidence Omission (= 1−CR)↓; **Dist** = Evidence Distortion↓; **Red** = Redundancy Waste↓; **UB** = Useful Budget↑; **AF** = candidate aggregation-failure rate: the percentage of all instances with a non-perfect answer and complete measured CR. MQ full-sharing AF is omitted because the saved 100−F1 values are not binary AF; see Table 8. Dist is omitted for summaries; Red and UB for full sharing. Darker pink = best per metric; lighter pink = 2nd; lighter blue = 3rd; darker blue = worst. AF is coverage-gated, so its rank alone does not measure protocol quality. Full breakdown for all ten models in Appendix C.2.2.

	Qwen-235B						Mistral-S						Qwen-3.6-F					
Protocol	CR	Om	Dist	Red	UB	AF%	CR	Om	Dist	Red	UB	AF%	CR	Om	Dist	Red	UB	AF%
Silo-Bench																		
Summary Exchange	.669	.331	—	.328	.379	53.8	.750	.250	—	.243	.502	63.5	.725	.275	—	.287	.477	50.0
Random Evidence Relay	.844	.156	.074	.167	.255	67.3	.885	.115	.030	.118	.287	67.3	.844	.156	.072	.139	.273	28.8
Score-Ranked Relay	.844	.156	.056	.167	.254	65.4	.885	.115	.030	.118	.286	69.2	.844	.156	.081	.139	.271	32.7
Redundancy-Aware Relay	.844	.156	.060	.167	.254	65.4	.885	.115	.032	.118	.286	69.2	.844	.156	.063	.139	.271	26.9
Full Evidence Sharing	1.00	.000	.000	—	—	76.9	1.00	.000	.000	—	—	82.7	1.00	.000	.000	—	—	34.6
MuSiQue																		
Summary Exchange	1.00	.000	—	.221	.776	79.0	1.00	.000	—	.159	.757	76.0	1.00	.000	—	.145	.874	65.0
Random Evidence Relay	1.00	.000	.030	.151	.319	72.0	1.00	.000	.045	.126	.322	79.0	1.00	.000	.030	.144	.360	53.0
Score-Ranked Relay	1.00	.000	.019	.151	.319	68.0	1.00	.000	.042	.126	.326	82.0	1.00	.000	.023	.144	.360	48.0
Redundancy-Aware Relay	1.00	.000	.023	.151	.319	72.0	1.00	.000	.049	.126	.326	79.0	1.00	.000	.023	.144	.360	51.0
Full Evidence Sharing	1.00	.000	.000	—	—	—	1.00	.000	.000	—	—	—	1.00	.000	.000	—	—	—

No-communication baselines. Three protocols require no interaction: (1) *Single Local* uses the first agent's answer; (2) *Majority Vote* takes the plurality of local answers; and (3) *Best-Local Oracle* selects a correct local answer when one exists, giving a no-communication upper bound.

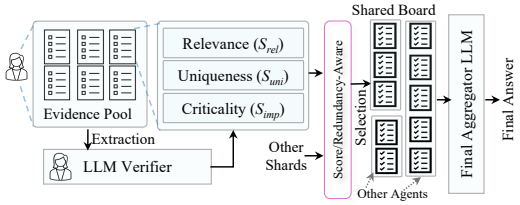


Figure 3: Illustration of score-ranked relay and redundancy-aware relay.

Natural-language communication. Four protocols transmit free-form or gated text: (4) *Free-form Debate* exchanges unconstrained arguments for one round (Du et al., 2024); (5) *Summary Exchange* sends one shard summary per agent; (6) *Confidence-Gated Communication* communicates only below a local-confidence threshold (Choi et al., 2025); and (7) *Disagreement-Gated Communication* communicates only when local answers disagree.

Structured evidence relay. Three protocols relay extracted atomic units (Figure 3): (8) *Random Evidence Relay* sends a uniformly random subset; (9) *Score-Ranked Evidence Relay* scores each unit for relevance, uniqueness, and local importance, then sends the top-k:

$$\begin{aligned} M_i &= \mathrm{TopK}_k(Z_i,\ s), \\ s(z) &= \tfrac{1}{3}(s_{\mathrm{rel}}(z) + s_{\mathrm{uniq}}(z) + s_{\mathrm{imp}}(z)); \end{aligned} \tag{12}$$

Protocol (10), *Redundancy-Aware Evidence Relay*, adds a diversity penalty:

$$M_i = \underset{\substack{M \subseteq Z_i \\ |M| \le k}}{\arg\max} \Big[\sum_{z \in M} s(z) \ - \ \rho \sum_{\substack{z, z' \in M \\ z \neq z'}} \mathrm{sim}(z, z') \Big]. \tag{13}$$

Pooled-evidence reference. (11) *Full Evidence Sharing* transmits every shard in full and serves as the pooled-evidence reference.

From diagnosis to repair. Extraction omission suggests re-extraction; selection omission suggests revised ranking or a targeted request; distortion suggests preserving source wording. With sufficient evidence, check integration and verification. Section 4.4 tests one model-placement choice; other repair policies remain untested, and gold supports are not assumed available at deployment.

Table 4: Budget k-sweep on Silo-Bench across five models. Each relay variant transmits up to k units per agent; extraction caps at $k_{\max}$=3. Score-Ranked beats Random for all five models at k=1 and four of five at k=2. At k=3, both variants transmit the capped extraction pool, so accuracy differences no longer isolate evidence selection. Acc (%); Q235B = Qwen-235B; Mis-S = Mistral-S; Q3.6F = Qwen-3.6-Flash; Gem-31B = Gemma-4-31B; Q80B-T = Qwen-80B-T.

	k=1					k=2					k=3				
Protocol	Q235B	Mis-S	Q3.6F	Gem-31B	Q80B-T	Q235B	Mis-S	Q3.6F	Gem-31B	Q80B-T	Q235B	Mis-S	Q3.6F	Gem-31B	Q80B-T
Random Relay	11.5	5.8	5.8	7.7	7.7	21.2	11.5	19.2	13.5	7.7	19.2	13.5	61.5	40.4	9.6
Score-Ranked Relay	15.4	9.6	57.7	25.0	9.6	17.3	15.4	57.7	36.5	11.5	19.2	17.3	53.8	40.4	21.2
Gap (Score−Rand)	*+3.9*	*+3.8*	*+51.9*	*+17.3*	*+1.9*	*−3.9*	*+3.9*	*+38.5*	*+23.0*	*+3.8*	*0.0*	*+3.8*	*−7.7*	*0.0*	*+11.6*

Table 5: Hop-count stratification on MuSiQue (Qwen-235B). F1 is mean answer F1 (%). AF is the percentage of all questions with a non-perfect answer despite complete measured critical-evidence coverage (relay rows). Full-sharing AF by hop is omitted because the saved $100 - \text{F1}$ values are not the same binary measure. "All hops" combines all 100 questions. Pink = lower AF or higher F1; blue = notably high AF ($\geq$77%). Multi-model results and exceptions in Appendix C.2.5; pooled binary AF comparison in Table 8.

	2-hop (n=50)		**3-hop** (n=35)		**4-hop** (n=15)		**All hops** (n=100)	
Protocol	F1	AF	F1	AF	F1	AF	F1	AF
Random Relay	51.1	62.0	34.9	82.9	33.9	80.0	42.9	72.0
Score-Ranked Relay	54.2	60.0	36.9	77.1	44.6	73.3	46.7	68.0
Redundancy-Aware Relay	50.4	66.0	37.4	80.0	36.9	73.3	43.9	72.0
Full Evidence Sharing	60.3	—	49.4	—	33.7	—	52.5	—

4 Experiments

We evaluate eleven protocols across ten LLMs on Silo-Bench (Zhang et al., 2026b), where agents coordinate over private numerical records, and distributed MuSiQue (Trivedi et al., 2022), where each agent holds one supporting passage for multi-hop reasoning. Task scores are exact-match accuracy on Silo-Bench and mean answer F1 on MuSiQue. Following the criticality study (Table 1), we compare systems, localize evidence losses, examine selection, and discuss repair targets. Model identifiers, full audit results, and the claim-to-evidence map appear in Appendices B.3, C.2.2, C.2.7, and C.1.

4.1 Audited Systems: Performance and Cost

Table 2 compares all protocols across both benchmarks. Full sharing generally performs strongly for capable models; atomic relay offers a lower-message-volume alternative, with performance varying by model and task. Free-form communication does not consistently outperform relay, while debate and gating show variable benefits across settings. The cost distinction matters for longer or repeated agent interactions, where exchanged context can accumulate, but fewer communication tokens do not imply fewer total inference tokens: on Silo-Bench with Qwen-235B, score-ranked relay exchanges 698 tokens versus 1,655 for full sharing while using 18,036 versus 6,089 total model tokens (Appendix C.2.1). Per-model performance and token breakdowns are in the same appendix.

Key Finding 1. Protocol accuracy varies by model and task. Relay can reduce exchanged tokens, but its extraction and scoring calls may raise total inference cost.

4.2 Attributing Evidence Loss and Residual Errors

On Silo-Bench, relay retains more measured critical evidence than summaries (Table 3). Table 6 localizes the larger loss to extraction: raw-to-extracted CritRec falls 7.5–20.0 points across four models, versus 3.7–6.0 points from extraction to transmission. All relay variants share an extraction pool, so selection cannot recover absent facts. Candidate aggregation failure still flags 26.9–71.2% of instances, calling for separate scrutiny of residual errors.

Table 6: Stage-wise critical evidence survival on Silo-Bench. *Raw* = 1.000 by construction. *Extracted* = CritRec over all extracted units before top-k selection. *Transmitted* = CritRec of selected/transmitted units. *AggFail* = candidate aggregation-failure rate: % of all instances with CritRec = 1 at transmission and a non-perfect answer. Darker pink = best per metric; lighter pink = 2nd; lighter blue = 3rd; darker blue = worst.

Model	Protocol	Raw	Extracted	Transmitted	AggFail (%)
Qwen-235B	Random Relay	1.000	0.881	0.844	67.3
	Score-Ranked Relay	1.000	0.881	0.844	65.4
	Redundancy-Aware Relay	1.000	0.881	0.844	65.4
Mistral-S	Random Relay	1.000	0.925	0.885	71.2
	Score-Ranked Relay	1.000	0.925	0.885	69.2
	Redundancy-Aware Relay	1.000	0.925	0.885	69.2
Qwen-80B	Random Relay	1.000	0.800	0.748	55.8
	Score-Ranked Relay	1.000	0.800	0.748	50.0
	Redundancy-Aware Relay	1.000	0.800	0.748	53.8
Qwen-3.6-F	Random Relay	1.000	0.904	0.844	28.8
	Score-Ranked Relay	1.000	0.904	0.844	32.7
	Redundancy-Aware Relay	1.000	0.904	0.844	26.9

On MuSiQue, relay often gives wrong answers despite complete measured coverage (Table 3). Scores generally decline with more hops, with exceptions (Table 5); its AF column flags non-perfect relay answers with complete measured coverage, and "all hops" combines the strata. Coverage is sensitive to lexical thresholds (Appendix C.4.1), so it does not establish intact transmission. Deleting a gold paragraph lowers scores by 12–37 points, supporting its relevance without validating each atomic-fact alignment (Appendix C.2.4).

Key Finding 2. On Silo-Bench, measured critical-fact loss is larger at extraction than selection. On MuSiQue, errors despite complete lexical coverage call for semantic and downstream checks, not a confirmed aggregation-failure diagnosis.

4.3 Local Importance Failure Analysis

Scoring quality. Figure 4 shows uniqueness predicting critical evidence better than relevance for several Silo-Bench models. At k=1, uniqueness-only selection retains 0.794 of critical units, versus 0.661 for the composite and 0.638 for relevance (Table 7); this does not establish universal optimality. On MuSiQue, local-importance and composite scores are near random (AUC $\approx$ 0.50; Appendix C.3.1), consistent with cross-shard dependencies.

Table 7: Score decomposition at k=1: critical recall (CritRec), false-negative rate (FNR), and mean scores of retained (TP) and dropped (FN) critical units. Gap = TP−FN.

Score component	CritRec	FNR	TP avg	FN avg	Gap
Relevance	0.638	0.373	1.000	1.000	0.000
Uniqueness	0.794	0.244	0.898	0.628	0.270
Importance	0.655	0.387	0.958	0.890	0.068
Overall (relay)	0.661	0.391	0.935	0.880	0.055

Key Finding 3. Under a one-unit Silo-Bench budget, uniqueness is the strongest tested local signal for retaining critical facts in the score decomposition; its advantage does not automatically transfer across models or tasks.

Budget sensitivity. Score-ranked relay beats random relay for all five Silo-Bench models at k=1 and four of five at k=2, averaging +13.1 points at k=2 (Table 4). Qwen-235B reverses by 3.9 points, and repeated-run uncertainty is unavailable. At k=3, both protocols send the capped extraction pool; remaining score differences may reflect ordering or generation variability, not content selection. The MuSiQue sweep shows a mean gain of 12.1 points at k=1 across five models (Appendix C.2.6). Uniqueness-only selection can also outperform the composite score (Appendix C.3.1), motivating scrutiny of how locally available signals approximate joint criticality.

Key Finding 4. Ranking helps at k=1 (five of five models) and k=2 (four of five); at k=3, both protocols send the same extracted facts.

Table 8: Audit-guided comparison of score-ranked relay and full evidence sharing for Qwen-235B. AggFail is the candidate aggregation-failure rate over all instances. Arrows indicate relay → full sharing; changes are in percentage points, calculated before rounding.

Data	N	CritRec	Task score	ΔScore	AggFail (%)	ΔAggFail
SB	52	0.844 → 1.000	19.2 → 23.1	+3.8	65.4 → 76.9	+11.5
MQ	100	1.000 → 1.000	46.7 → 52.5	+5.8	68.0 → 65.0	−3.0

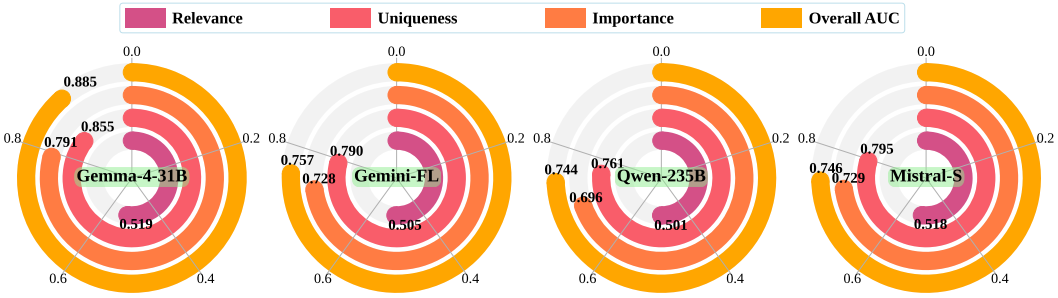


Figure 4: Local importance failure: AUC for predicting gold criticality from each score component (Silo-Bench). Each ring represents one score component; bar length indicates AUC (random = 0.5, full ring = 1.0). Full Spearman/P@3/R@3 per score in Appendix C.3.1.

4.4 From Attribution to Repair Targets

Table 8 compares relay with full sharing on matched Qwen-235B instances. Full sharing raises task scores but changes evidence quantity and representation; its higher Silo-Bench candidate aggregation-failure rate (AggFail) reflects coverage gating.

We test a more targeted choice on 326 score-ranked relay failures across Gemma-4-31B, GPT-4o-mini, and Qwen-235B. Each selected answer is non-perfect despite complete measured critical-fact coverage, so the audit flags a *candidate* aggregation gap. Both arms start from the same saved board. One Gemini 3.6 Flash call either produces the final answer from that board or replaces one predetermined agent's extracted evidence before the baseline model aggregates the revised board. Table 9 reports the saved original scores and equal-weight task means across models (MuSiQue F1; Silo-Bench exact match); N counts selected cases.

Table 9: One-call model placement on flagged failures (0–100). MQ = MuSiQue; SB-5/10 = Silo-Bench agent counts; Agg./Ext. = Gemini aggregation/extraction; Δ = Agg. − Ext.

Task	N	Orig.	Agg.	Ext.	Δ
MQ	213	15.8	28.5	22.4	+6.1
SB-5	56	0.0	32.6	6.0	+26.6
SB-10	57	0.0	10.2	5.8	+4.4

At the same Gemini-call budget, final aggregation beats local extraction by 6.1, 26.6, and 4.4 points across the three tasks. It also improves every task mean relative to the original run. This supports spending a fixed intervention budget at the audit-flagged stage. Lexical candidate labels and unequal inputs (whole board versus one shard) limit causal attribution; see Appendix C.4.4. The five-agent Silo-Bench setting shows the largest gap: 32.6 points for final aggregation versus 6.0 for local extraction. The arms match Gemini calls, not total inference tokens. These task averages support placing the stronger call at the flagged stage; they do not establish every case's causal bottleneck.

5 Conclusion

The evidence-flow audit separates extraction and selection losses, transmission distortion, and residual errors across eleven protocols, two benchmarks, and ten LLMs. On flagged failures, one stronger-model call yields larger gains at final aggregation than at local extraction across three tasks. This supports audit-guided budget allocation, while semantic validation and broader repair controls remain necessary (Appendix D).

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and editing, LaTeX preparation, and interpretation of the reported experimental results. The LLMs evaluated as agents and used for automated judgments are described in the methods and appendix. The authors reviewed AI-assisted material against the underlying results and take responsibility for the final text, analyses, and claims.

ETHICS STATEMENT

This study evaluates LLM behavior on existing Silo-Bench and MuSiQue benchmark tasks and did not conduct a new human-subject study. Automated evidence-coverage and error labels are imperfect proxies; we report their limitations and do not treat them as definitive assessments of reasoning quality. The findings should not be assumed to establish reliability in consequential applications without further validation.

REPRODUCIBILITY STATEMENT

Appendices B and C.1 document the benchmark setup, model identifiers, protocols, prompts, inference settings, audit metrics, and sample sizes. The extended results and sensitivity analyses appear in Appendices C.2 and C.4. Experiment and table-generation scripts are maintained in the project codebase.

REFERENCES

Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei A Zaharia, Joseph E Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems*, 2025.

Kaiyan Chang, Yonghao Shi, Chenglong Wang, Hang Zhou, Chi Hu, Xiaoqian Liu, Yingfeng Luo, Yuan Ge, Tong Xiao, and Jingbo Zhu. Step-level verifier-guided hybrid test-time scaling for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18473–18488, 2025.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? *arXiv preprint arXiv:2508.17536*, 2025.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*, 2024.

Wei Fan, JinYi Yoon, and Bo Ji. imad: Intelligent multi-agent debate for efficient and accurate llm inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29403–29411, 2026.

Adit Jain, Vikram Krishnamurthy, and Yiming Zhang. Collaborative QA using interacting LLMs: Impact of network structure, node capability and distributed data. *arXiv preprint arXiv:2511.14098*, 2025.

Harsh Jhamtani, Jacob Andreas, and Benjamin Van Durme. Llm agents for coordinating multi-user information gathering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17800–17826, 2025.

Can Jin, Hongwu Peng, Qixin Zhang, Yujin Tang, Dimitris N Metaxas, and Tong Che. Two heads are better than one: Test-time scaling of multi-agent collaborative reasoning. *arXiv preprint arXiv:2504.09772*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Yuxuan Li, Aoi Naito, and Hirokazu Shirado. Hiddenbench: Assessing collective reasoning in multi-agent llms via hidden profile tasks. *arXiv preprint arXiv:2505.11556*, 2025.

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, 2024.

Shalev Lifshitz, Sheila A McIlraith, and Yilun Du. Multi-agent verification: Scaling test-time compute with multiple verifiers. *arXiv preprint arXiv:2502.20379*, 2025.

Yuhan Liu, Juntian Zhang, Yichen Wu, Martin Takac, Salem Lahlou, Xiuying Chen, and Nils Lukas. Breaking the martingale curse: Multi-agent debate via asymmetric cognitive potential energy. *arXiv preprint arXiv:2603.06801*, 2026.

Jacob Marschak and Roy Radner. *Economic Theory of Teams*. Yale University Press, New Haven, CT, 1972.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

Sunil Prakash. From debate to deliberation: Structured collective reasoning with typed epistemic acts. *arXiv preprint arXiv:2603.11781*, 2026.

Roy Radner. Team decision problems. *The Annals of Mathematical Statistics*, 33(3):857–881, 1962.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Xu Shen, Yixin Liu, Yiwei Dai, Yili Wang, Rui Miao, Yue Tan, Shirui Pan, and Xin Wang. Understanding the information propagation effects of communication topologies in LLM-based multi-agent systems. *arXiv preprint arXiv:2505.23352*, 2025.

Jinwei Su, Yinghui Xia, Qizhen Lan, Xinyuan Song, Yang Jingsong, Lewei He, and Tianyu Shi. Difficulty-aware agent orchestration in llm-powered workflows. *arXiv e-prints*, pp. arXiv–2509, 2025.

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.

Jingbo Wang, Sendong Zhao, Jiatong Liu, Haochun Wang, Wanting Li, Bing Qin, and Ting Liu. Orchestrating intelligence: Confidence-aware routing for efficient multi-agent collaboration across multi-scale models. *arXiv preprint arXiv:2601.04861*, 2026.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*, 2025.

Yongjin Yang, Euiin Yi, Jongwoo Ko, Kimin Lee, Zhijing Jin, and Se-Young Yun. Revisiting multi-agent debate as test-time scaling: A systematic study of conditional effectiveness. *arXiv preprint arXiv:2505.22960*, 2025.

Meiru Zhang, Zaiqiao Meng, and Nigel Collier. Failure modes in multi-hop QA: The weakest link effect and the recognition bottleneck. *arXiv preprint arXiv:2601.12499*, 2026a.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, and Qingyun Wu. Which agent causes task failures and when? on automated failure attribution of LLM multi-agent systems. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76583–76599. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25cq.html.

Yuzhe Zhang, Feiran Liu, Yi Shan, Xinyi Huang, Xin Yang, Yueqi Zhu, Xuxin Cheng, Cao Liu, Ke Zeng, Terry Jingchen Zhang, et al. Silo-bench: A scalable environment for evaluating distributed coordination in multi-agent llm systems. *arXiv preprint arXiv:2603.01045*, 2026b.

Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. Demystifying multi-agent debate: The role of confidence and diversity. *arXiv preprint arXiv:2601.19921*, 2026.

APPENDIX CONTENTS

A. Related Work
- A.1 Multi-Agent Communication and Debate
- A.2 Test-Time Scaling and Verification
- A.3 Distributed Evidence Benchmarks and Communication Analysis

B. Implementation Details
- B.1 Problem Setting
- B.2 Benchmark Selection and Rationale
- B.3 Models
- B.4 Protocol Implementation
- B.5 Gold Critical Unit Approximation
- B.6 Metric Computation
- B.7 Prompt Details

C. Experiments and Additional Analyses
- C.1 Experiment Setup
- C.2 Extended Results
  - C.2.1 Per-Model Full Results
  - C.2.2 Multi-Model Evidence-Flow Audit
  - C.2.3 Stage-Wise Evidence Survival
  - C.2.4 Critical-Evidence Bottleneck
  - C.2.5 Hop-Count Stratification on MuSiQue
  - C.2.6 Budget k-Sweep on MuSiQue
  - C.2.7 Additional Model Robustness Evaluation
- C.3 Additional Analyses
  - C.3.1 Local Scoring and Joint-Criticality Analysis
  - C.3.2 Evidence Fidelity: Audit Details and Model Comparisons
- C.4 Measurement and Interaction Controls
  - C.4.1 Coverage-Threshold Sensitivity
  - C.4.2 Zero-Evidence Controls
  - C.4.3 Fixed-Evidence-Budget Multi-Round Pilot
  - C.4.4 Residual Error Categories and Judge Validation

D. Limitations and Future Work

A Related Work

We organize the related work into three themes that bracket this paper: multi-agent communication protocols, test-time scaling and verification, and distributed evidence benchmarks with failure analysis. A structured survey table is provided at the end of this section.

A.1 Multi-Agent Communication and Debate

A central line of work treats agent interaction as iterative critique and refinement. Du et al. (2024) show that repeated rounds of challenge and rebuttal improve factuality in factual QA tasks; Wang et al. (2024) extend this with layered aggregation across heterogeneous models (Mixture-of-Agents), where outputs from one layer serve as proposals for the next. Jin et al. (2025) introduce a CEO-agent variant that adaptively controls discussion depth based on the difficulty of the current instance.

These debate-based systems raise the question of when communication helps. Yang et al. (2025) demonstrate that debate gains are conditional on task type and on whether the correct answer is held by a majority of agents. Subsequent failure-mode analyses reveal a complementary risk: free-form exchange can suppress minority-held correct answers when evidence is asymmetrically distributed. Liu et al. (2026) propose asymmetric cognitive potential energy to rescue minority-truth, while Prakash (2026) introduce typed deliberative acts with minority report preservation. Choi et al. (2025); Wynn et al. (2025); Zhu et al. (2026) further characterize conditions under which multi-agent debate under- or over-performs single-agent baselines.

Selective gating methods reduce overhead by triggering communication only when needed. Confidence-based gating (Fan et al., 2026) predicts whether debate will improve accuracy before committing to it; difficulty-aware orchestration (Su et al., 2025) routes instances to different model scales based on estimated complexity; confidence-aware routing (Wang et al., 2026) selects which agents to activate at each step. These methods treat the *existence* of communication as a design variable but leave its *content* free-form. The evidence-flow audit addresses a complementary question: given that agents communicate, where in the pipeline is evidence lost, and which failure mode is dominant for a given task type?

A.2 Multi-Agent Failure Attribution

Two recent works directly study diagnosing failures in LLM multi-agent systems. Cemri et al. (2025) introduce MAST, a taxonomy of 14 failure modes grouped into system-design issues, inter-agent misalignment, and task verification, and use it to analyze execution traces with an LLM-as-a-judge pipeline. Zhang et al. (2025) formulate automated failure attribution as identifying both the agent responsible for a failed task and the decisive error step; their Who&When benchmark contains annotated failure logs from 127 multi-agent systems. These frameworks provide system-level categories and agent/step-level blame attribution, respectively. Our audit instead follows critical evidence through the pipeline, localizing where evidence is omitted or distorted and flagging cases where available evidence was not used successfully. The analyses are complementary: evidence-flow traces can offer a more targeted account of information loss, while MAST and Who&When cover broader execution failures and responsibility assignment.

A.3 Test-Time Scaling and Verification

A parallel trend scales test-time compute through verifier-based evaluation of candidate reasoning steps. Lifshitz et al. (2025) show that using multiple verifiers to score answer candidates improves performance on reasoning benchmarks as verifier count increases. Chang et al. (2025) propose step-level verifier-guided hybrid scaling, where intermediate reasoning steps are evaluated and pruned before final answer generation. These approaches share with our evidence relay the idea of assigning a quality score to generated content before it is used downstream. Score-ranked evidence relay borrows the verifier pattern for evidence scoring: instead of certifying a candidate answer, the scoring step rates each extracted atomic fact on relevance, uniqueness, and estimated local importance. Our experiments show, however, that this local scoring is unreliable because evidence criticality is joint across shards—a fact that appears unimportant within one agent's shard may become decisive only when combined with evidence held by another agent.

Factual consistency evaluation (Min et al., 2023) provides the methodological precedent for atomizing claims into independently verifiable units; we adapt this decomposition to the distributed setting to make evidence units trackable through each pipeline stage. Retrieval-augmented generation (Lewis et al., 2020) is the centralized analogue: a single retriever surfaces relevant passages from a shared corpus, whereas our setting has no shared retriever and each agent holds only its private shard.

A.4 Distributed Evidence Benchmarks and Communication Analysis

Our evaluation uses two benchmarks designed for distributed evidence settings. Zhang et al. (2026b) introduce Silo-Bench for numerical coordination tasks where agents hold non-overlapping private records and must collectively produce an exact answer; Trivedi et al. (2022) provide MuSiQue, which we adapt to a distributed variant by assigning each supporting paragraph to a separate agent. Li et al. (2025) introduce HiddenBench for hidden-profile tasks, where correct collective reasoning requires surfacing asymmetric private evidence that individual agents would not prioritize—closely related to the omission failure mode we study.

From an information-theoretic perspective, team theory (Marschak & Radner, 1972; Radner, 1962) provides the formal grounding: a group of agents with a shared payoff who observe only local signals must design a communication policy that maximizes joint outcome at bounded cost. The evidence-flow audit operationalizes this by tracking whether the information with the highest marginal value for the joint answer survives the communication channel.

These distributed inference-time settings are distinct from gradient-aggregation methods such as federated learning (McMahan et al., 2017): model parameters are fixed and agents coordinate at inference time by transmitting compact evidence representations, not parameter updates. The critical difference is that mistakes are not correctable across rounds—each inference pass is final, making evidence selection quality immediately consequential.

Table 10: Related works grouped by category. Our work draws from multi-agent communication, team theory, and evidence-grounded reasoning.

Category	Method	Title	Core idea
Debate / TTS	MAD (Du et al., 2024)	Improving Factuality and Reasoning in Language Models through Multiagent Debate	Iterative multi-agent critique–rebuttal debate.
Debate / TTS	MoA (Wang et al., 2024)	Mixture-of-Agents Enhances Large Language Model Capabilities	Layered agent aggregation and refinement.
Debate / TTS	CEO-agent (Jin et al., 2025)	Two Heads are Better Than One: Test-time Scaling of Multi-agent Collaborative Reasoning	CEO-guided adaptive discussion and reasoning depth.
Debate Analysis	MAD-as-TTS (Yang et al., 2025)	Revisiting Multi-Agent Debate as Test-Time Scaling	Conditional effectiveness analysis of debate scaling.
Debate Gating	iMAD (Fan et al., 2026)	iMAD: Intelligent Multi-Agent Debate for Efficient and Accurate LLM Inference	Debate triggering via lightweight benefit prediction.
Workflow Orchestration	DAAO (Su et al., 2025)	Difficulty-Aware Agent Orchestration in LLM-Powered Workflows	Difficulty-aware workflow, operator, and model orchestration.
Routing / MAS	OI-MAS (Wang et al., 2026)	Confidence-Aware Routing for Efficient Multi-Agent Collaboration	Confidence-aware routing of roles and model scales.
Verification / TTS	MAV (Lifshitz et al., 2025)	Multi-Agent Verification: Scaling Test-Time Compute with Multiple Verifiers	Verifier-count scaling for candidate evaluation.
Verification / TTS	Hybrid TTS (Chang et al., 2025)	Step-level Verifier-guided Hybrid Test-Time Scaling for Large Language Models	Step-level verifier-guided hybrid scaling.
Collective Reasoning	HiddenBench (Li et al., 2025)	HiddenBench: Assessing Collective Reasoning in Multi-Agent LLMs via Hidden Profile Tasks	Hidden-profile benchmark for asymmetric evidence integration.
Distributed Coordination	Silo-Bench (Zhang et al., 2026b)	Silo-Bench: A Scalable Environment for Evaluating Distributed Coordination	Scalable benchmark with private, non-overlapping agent evidence.
Federated Learning	FedAvg (McMahan et al., 2017)	Communication-Efficient Learning of Deep Networks from Decentralized Data	Federated optimization via gradient aggregation.
Team Theory	Marschak & Radner (1972)	Economic Theory of Teams	Formal treatment of shared-payoff decentralized-information teams.
Team Theory	Radner (1962)	Team Decision Problems	Information structure and optimal communication in teams.
RAG	RAG (Lewis et al., 2020)	Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks	Centralized retrieval over document collections for open-domain QA.
Minority-Truth Rescue	AceMAD (Liu et al., 2026)	Breaking the Martingale Curse: Multi-Agent Debate via Asymmetric Cognitive Potential Energy	Minority-truth amplification under correlated errors.
Deliberation Protocol	DCI (Prakash, 2026)	Structured Collective Reasoning with Typed Epistemic Acts	Typed deliberative acts with minority report preservation.

B Implementation Details

B.1 Problem Setting

A distributed reasoning instance is a tuple $x = (q, \{E_i\}_{i=1}^n, y, S^\star)$, where q is the query, E_i is the private evidence shard observed by agent i, y is the gold answer, and $S^\star$ is the set of gold critical evidence units—a minimal set of atomic facts sufficient to derive y.

Each agent extracts atomic evidence units from its shard, $Z_i = \mathrm{Extract}(q, E_i)$, and a communication protocol selects a subset to transmit, $M_i = \pi_i(q, Z_i)$. All transmitted messages are pooled into a shared evidence board $\mathcal{B} = \bigcup_{i=1}^n M_i$, and a final aggregator produces the answer $\hat{y} = A(q, \mathcal{B})$.

An **evidence unit** is an atomic factual claim: an entity–relation–entity triple, a numerical value, a temporal constraint, or a key-value observation—minimal and self-contained, independently interpretable without surrounding context. A **critical evidence unit** $z \in S^\star$ is one whose removal changes answerability or substantially reduces the probability of deriving the correct answer.

The audit decomposes failures across five pipeline stages: sharding, extraction, transmission, consolidation, and aggregation. The metrics (CritRecall, OmissionRate, DistortionRate, RedundancyRate/UsefulBudget, AggFail) support attribution across checkpoints rather than a one-to-one assignment of every metric to a stage (§3). Final-board omission alone does not identify its origin.

B.2 Benchmark Selection and Rationale

We select two benchmarks that stress-test qualitatively different stages of the evidence pipeline.

Silo-Bench (Zhang et al., 2026b). Silo-Bench is a synthetic benchmark for distributed numerical coordination. Each task presents a question involving a precise numerical answer derivable only by combining private records held by different agents (e.g., a regional total computed from per-agent observations). Agents hold non-overlapping shards: no agent can answer correctly from its own shard alone. This design makes the benchmark useful for studying evidence extraction and selection: the relevant numerical values must reach the aggregator, which must also combine them correctly. The benchmark scales across two agent-count tiers (n=5 and n=10), enabling analysis of how failure modes change with team size. We use both tiers in all experiments to report results over the combined set of tasks.

Distributed MuSiQue (Trivedi et al., 2022). MuSiQue is a multi-hop factual QA dataset in which each question requires chaining 2–4 supporting paragraphs. We adapt it to the distributed setting by assigning each supporting paragraph to a separate agent as its private shard; agents with non-supporting paragraphs hold distractors. This design tests cross-shard evidence integration: supporting paragraphs provide the facts needed for multi-step inference, but extracted units need not preserve those facts completely. Using MuSiQue allows us to compare measured evidence loss and residual answer errors across task structures without assuming a dominant failure mode in advance.

B.3 Models

We evaluate ten models that together span a wide range of model series, providers, architectures, and capability tiers. The selection was guided by two practical criteria: (1) diversity across publicly available model families to avoid conclusions tied to any single provider or design choice, and (2) cost-effective API access via OpenRouter, which allowed us to run large-scale multi-agent experiments across all models without prohibitive compute costs. The resulting set covers open-weight and proprietary models, dense and mixture-of-experts architectures, standard instruction-tuned and chain-of-thought-optimized variants, and sizes ranging from 3.6 B to over 1 T parameters. Table 11 lists the short names used throughout the paper and the full model identifiers used via the OpenRouter API.

Qwen-235B (qwen/qwen3-235b-a22b-2507). Qwen-235B is a large mixture-of-experts model from Alibaba, activating 22 B parameters per forward pass from a 235 B total. It represents the highest-capability end of our model set and serves as the primary model for detailed analyses (evidence distortion, joint criticality) due to its strong instruction-following and structured-output reliability.

Table 11: Short names and full model identifiers for all ten models.

Short name	Full model ID
Qwen-235B	`qwen/qwen3-235b-a22b-2507`
Mistral-S	`mistralai/mistral-small-2603`
GPT-OSS	`openai/gpt-oss-safeguard-20b`
Qwen-80B	`qwen/qwen3-next-80b-a3b-thinking`
Ring-1T	`inclusionai/ring-2.6-1t`
Qwen-3.6-F	`qwen/qwen3.6-flash`
Gemini-FL	`google/gemini-2.5-flash-lite`
GPT-4o-m	`openai/gpt-4o-mini`
DeepSeek-32B	`deepseek/deepseek-r1-distill-qwen-32b`
Gemma-4-31B	`google/gemma-4-31b-it`

Mistral-S (mistralai/mistral-small-2603). Mistral-Small is a mid-size dense model from Mistral AI, designed for efficiency while maintaining competitive reasoning quality. It provides a mid-range reference point from the Mistral model family.

GPT-OSS (openai/gpt-oss-safeguard-20b). GPT-OSS is a 20 B open-source safety-focused model released by OpenAI. Its inclusion covers the OpenAI model family at a smaller, open-weight scale distinct from the proprietary GPT-4o series.

Qwen-80B-T (qwen/qwen3-next-80b-a3b-thinking). Qwen-80B-T is a thinking-optimized model in the Qwen3 family that produces extended chain-of-thought reasoning before its final answer. It represents the chain-of-thought/reasoning-model tier and tests whether explicit intermediate reasoning steps affect evidence-flow behavior.

Ring-1T (inclusionai/ring-2.6-1t). Ring-1T is an extremely large model with approximately 1 T parameters, designed with long-context handling as a primary capability. It tests whether scale alone—independent of other design choices—affects evidence omission and aggregation patterns.

Qwen-3.6-F (qwen/qwen3.6-flash). Qwen-3.6-Flash is a lightweight flash variant from the Qwen family, balancing low cost with reasonable capability. It serves as the efficiency-tier representative and captures the failure-mode profile of smaller, cheaper models.

Gemini-FL (google/gemini-2.5-flash-lite). Gemini-2.5-Flash-Lite is Google's lightweight API model from the Gemini 2.5 generation. It provides coverage of the Google model family at the efficient tier.

GPT-4o-m (openai/gpt-4o-mini). GPT-4o-mini is OpenAI's cost-efficient small model from the GPT-4o family. It covers the OpenAI family at a larger scale than GPT-OSS and represents the widely-used commercial small-model tier.

DeepSeek-32B (deepseek/deepseek-r1-distill-qwen-32b). DeepSeek-R1-Distill-Qwen-32B is a 32 B model obtained by distilling DeepSeek-R1's reasoning capabilities into a Qwen-based dense model. It represents the reasoning-distillation paradigm and also serves as a case study in degenerate extraction behavior (§C.2.2).

Gemma-4-31B (google/gemma-4-31b-it). Gemma-4-31B is Google's open-weight 31 B instruction-tuned model from the Gemma 4 family. It covers the Google open-weight tier and complements Gemini-FL to separate the effects of open vs. proprietary deployment from model capability.

All ten models are used for the main protocol comparison (Table 2); the subset with complete extraction data is used for the full evidence-flow audit.

Algorithm 1 Score-Ranked Evidence Relay (with optional redundancy penalty)

Input: query q, private shards $\{E_i\}_{i=1}^{n}$, per-agent budgets $\{k_i\}$, redundancy penalty $\rho \geq 0$
Output: final answer $\hat{y}$
for each agent i *(in parallel)* **do**
  $Z_i \leftarrow \textsc{ExtractUnits}(q,\, E_i)$
  for each $z \in Z_i$ **do**
    $s_{\text{rel}},\, s_{\text{uniq}},\, s_{\text{imp}} \leftarrow \textsc{LocalImportanceScore}(q,\, z)$
    $s(z) \leftarrow (s_{\text{rel}} + s_{\text{uniq}} + s_{\text{imp}})/3$
  end for
  if $\rho = 0$ **then**
    $M_i \leftarrow \textsc{TopK}(Z_i,\, s,\, k_i)$
  else
    $M_i \leftarrow \textsc{RedundancySelect}(Z_i,\, s,\, k_i,\, \rho)$ (Eq. 13)
  end if
end for
$\mathcal{B} \leftarrow \bigcup_{i=1}^{n} M_i$
$\hat{y} \leftarrow \textsc{Aggregate}(q,\, \mathcal{B})$
return $\hat{y}$

B.4 Protocol Implementation

Algorithm 1 describes the core evidence relay pipeline, which covers both Score-Ranked and Redundancy-Aware relay depending on whether the diversity penalty ρ is non-zero. All relay protocols share the same extraction and aggregation steps; they differ only in the selection step (random vs. score-ranked vs. redundancy-penalized).

The extraction step calls each agent's local LLM to decompose its private shard into atomic evidence units. Each extracted unit is then scored independently on four dimensions: relevance, uniqueness, local importance, and joint dependency (see Prompt Details, §B.7). For Redundancy-Aware relay, a greedy submodular selection is applied to maximize total score subject to a diversity penalty on pairwise semantic similarity between selected units. The aggregation step presents all transmitted units to the aggregator in a fixed format and requires structured output (ANSWER/USED_EVIDENCE/RATIONALE).

B.5 Gold Critical Unit Approximation

We approximate the gold critical unit set $S^{\star}$ from benchmark metadata. For MuSiQue, we treat each paragraph flagged as `is_supporting=True` as one critical evidence unit. For Silo-Bench, we use the `gold_supports` or `supporting_facts` fields where available.

This approximation has two known limitations. First, some tasks may have multiple valid minimal support sets; we use all annotated gold paragraphs and may over-count critical units. Second, the granularity of gold paragraphs is coarser than atomic evidence units; a gold paragraph may contain one critical fact and several non-critical sentences.

In the bottleneck intervention experiments, we remove gold paragraphs (not individual sentences) to ensure that the critical fact is actually removed. In the alignment step, we compare transmitted atomic units to the full text of gold paragraphs using lexical similarity.

B.6 Metric Computation

This subsection explains how each metric is computed and the rationale behind its design. Metrics fall into two groups: *audit metrics* that measure evidence flow through the pipeline, and *result metrics* that measure end-task performance.

Audit Metrics

Critical Recall (CritRec). CritRec measures the fraction of gold critical units that survive transmission and are present in the set of units available to the aggregator. A transmitted unit t is said

Algorithm 2 CritRecall via lexical alignment

Require: Gold units $S^\star = \{g_1, \ldots, g_m\}$, transmitted units $T = \{t_1, \ldots, t_k\}$, threshold $\tau = 0$

```
covered ← 0
for each g ∈ S* do
    s_max ← max_{t∈T} s(g, t)
    if s_max > τ then
        covered ← covered + 1
    end if
end for
return CritRec = covered / |S*|
```

to *cover* gold unit g if their lexical similarity exceeds a threshold τ. The implemented matcher uses $s(g,t) = \max\{J(g,t), N(g,t)\}$, where J is Jaccard overlap over lowercase whitespace-token sets and N is the fraction of numeric strings in g appearing in t (zero if g has no numeric strings). Coverage requires $s(g,t) > \tau$, with default $\tau = 0$. This permissive rule measures lexical overlap rather than semantic sufficiency; Appendix C.4.1 reports sensitivity to stricter thresholds. Algorithm 2 states the computation formally.

Omission Rate (Om). OmissionRate $= 1 -$ CritRec. It directly measures the fraction of critical evidence lost before the aggregator sees it. Comparing extraction and transmission checkpoints is necessary to separate extraction omission from selection omission.

Distortion Rate (Di). Among all transmitted units that are aligned to at least one gold critical unit (i.e., $s_{\max} > \tau$), DistortionRate is the fraction whose content is semantically altered relative to the gold unit. We detect distortion via an LLM judge that is given the original gold text and the transmitted paraphrase and asked whether any factual content—especially numerical values—has changed. DistortionRate measures the *quality* of evidence that does reach the aggregator; it is orthogonal to CritRec.

Redundancy Rate (Re). RedundancyRate is the mean pairwise Jaccard similarity among transmitted units, using lowercase whitespace-token sets in the audit script. It is zero when fewer than two units are transmitted. This is a lexical similarity diagnostic, not a thresholded duplicate fraction or a direct measure of lost critical facts.

Useful Budget (UB). UsefulBudget divides the number of distinct transmitted texts selected as best matches for covered gold supports by the number of transmitted units:

$$\text{UB} = \frac{|\text{distinct best-match texts for covered supports}|}{|T|}.$$

It is zero for an empty board and is not the complement of RedundancyRate. The implementation deduplicates matched texts in the numerator; the measure concerns evidence units, not token cost or verified semantic utility.

Candidate Aggregation Failure (AggFail). Operational AggFail flags tasks with complete measured lexical coverage (CritRec $= 1$) and a non-perfect answer, making them candidates for aggregation-failure analysis. Such cases motivate checking multi-hop reasoning, numerical combination, and answer extraction, while allowing for false-positive lexical coverage. The reported percentage uses all evaluated instances as its denominator; distortion is audited separately and is not an additional gate in the implemented AF calculation. Algorithm 3 states the per-task classification and dataset-level aggregation.

AggFail is not a certified bound on true aggregation failure: lexical matching can yield both false-positive and false-negative coverage, and answer matching can reject a semantically correct response. Semantic evidence validation and answer checking are needed before assigning a causal interpretation.

Algorithm 3 AggFail computation

Require: Per-task CritRec scores $\{r_i\}$, per-task scores $\{s_i\}$ on a 0–100 scale, tasks $\mathcal{D}$

$F \leftarrow 0$
for each task $i \in \mathcal{D}$ **do**
  if $r_i = 1.0$ **then**
    if $s_i < 100$ **then**
      $F \leftarrow F + 1$
    end if
  end if
end for
return AggFail% $= 100F/|\mathcal{D}|$

Result Metrics

Accuracy (Acc). The task score is exact-match accuracy on Silo-Bench and mean normalized token-level answer F1 on MuSiQue. Binary success for the supplementary failure analysis requires answer F1 equal to 1; mean F1 and the percentage of perfect answers are distinct quantities. For Silo-Bench, we normalize by stripping currency symbols, commas, and trailing zeros before comparing. For MuSiQue, we follow the standard SQuAD-style answer normalization (lowercasing, removing articles and punctuation).

Communication Tokens (CommTok). CommTok is the total number of tokens exchanged among agents and between agents and the aggregator during one task instance. We count only the content of inter-agent messages, not the system prompts or internal chain-of-thought tokens. Lower CommTok indicates higher communication efficiency.

Token Cost (TokCost). TokCost is the mean total LLM API tokens consumed across extraction, scoring, and aggregation for an instance, including tokens outside inter-agent messages.

Utility. Utility is the reported scalar $\text{Acc}/100 - 10^{-4} \cdot \text{CommTok}$, where Acc is percentage accuracy. It trades task accuracy against exchanged tokens in the appendix tables, but does not account for total inference tokens.

Recovery. Recovery is accuracy under the Oracle protocol (full evidence broadcast to all agents) minus accuracy under the evaluated protocol, expressed as a percentage of the oracle gap. It answers "how much of the theoretical ceiling did the protocol recover?" and controls for task difficulty when comparing protocols across benchmarks.

B.7 Prompt Details

This section reproduces all prompt templates used in the evidence-flow audit pipeline. Each LLM call is issued with a system instruction and a user prompt. The local importance scoring call uses temperature 0.0 for reproducibility; all other generation calls use temperature 0.7.

Evidence extraction — system instruction.

```
You are an evidence extraction assistant.
Extract atomic factual units from the provided evidence that may help
answer the question. Each unit must be:
  - self-contained (interpretable without other context)
  - specific (include exact names, numbers, dates)
  - factual (not a reasoning step or answer guess)
Output a numbered list of atomic evidence units, one per line.
Do not add explanations.
```

Evidence extraction — user prompt.

```
Question: {question}

Evidence shard:
{context}

Extract up to {max_units} atomic evidence units from the shard above.
Output a numbered list (one unit per line).
```

Local importance scoring — system instruction.

```
You are scoring whether an evidence unit should be communicated in a
distributed reasoning system.

Return JSON only with no extra text:
{"relevance": <0-1>, "uniqueness": <0-1>, "local_importance": <0-1>,
 "joint_dependency": <0-1>, "overall": <0-1>}
```

Local importance scoring — user prompt.

```
Question: {question}
Evidence unit: {unit}

Score from 0 to 1:
- relevance: Does the unit relate to the question?
- uniqueness: Is the unit likely unavailable to other agents?
- local_importance: Could omitting this unit change the answer?
- joint_dependency: Could this unit become important only when combined
  with evidence from other agents?

Return JSON only:
{"relevance": ..., "uniqueness": ..., "local_importance": ...,
 "joint_dependency": ..., "overall": ...}
```

Alignment judge — system instruction.

```
You are an evidence alignment judge.

Given a gold evidence unit and a candidate transmitted unit, classify
their relationship.

Labels:
- EXACT: candidate preserves the same meaning and all key entities,
  numbers, dates, and constraints.
- PARAPHRASE: candidate preserves the same meaning with different
    wording.
- PARTIAL: candidate preserves some information but omits a key
    constraint.
- DISTORTED: candidate is related but changes a key entity, number,
  date, relation, negation, or condition.
- NO_MATCH: candidate does not express the gold evidence.

Return JSON only:
{"label": "EXACT|PARAPHRASE|PARTIAL|DISTORTED|NO_MATCH",
 "confidence": 0.0, "reason": "short explanation"}
```

Alignment judge — user prompt.

```
Gold unit: {gold_unit}
Candidate unit: {candidate_unit}
```

```
Classify the relationship.
Return JSON only:
{"label": "...", "confidence": ..., "reason": "..."}
```

Aggregation — system instruction.

```
You are a question-answering assistant. The following atomic evidence
units were transmitted from private evidence shards. Use ONLY these
units to answer the question.
Output:
ANSWER: <your concise answer>
USED_EVIDENCE: <comma-separated unit numbers you relied on>
RATIONALE: <one sentence explaining how the evidence supports the answer
    >
If the units are insufficient, write:
ANSWER: Insufficient information.
```

Aggregation — user prompt.

```
Transmitted evidence units:
{numbered_evidence_list}

Question: {question}

Answer using ONLY the transmitted units above.
ANSWER:
USED_EVIDENCE:
RATIONALE:
```

Summary exchange — agent system instruction.

```
You are a summarization assistant in a multi-agent reasoning system.
Given a private evidence shard and a question, produce a concise summary
of the key facts in your shard that may be relevant to answering the
question. Focus on factual content; do not answer the question.
Be as concise as possible.
```

Summary exchange — agent user prompt.

```
Your private evidence:
{context}

Question: {question}

Write a concise summary (at most {max_words} words) of the key facts
in your evidence that are relevant to this question.
Do not answer the question directly.
```

Summary exchange — aggregation system instruction.

```
You are a question-answering assistant. You have received concise
summary reports from multiple agents, each with access to different
private evidence. Use only the information in these summaries ---
do not use outside knowledge.
Output your final answer as:
ANSWER: <your concise answer>
```

Summary exchange — aggregation user prompt.

```
Agent 1 summary:
{summary_1}

Agent 2 summary:
{summary_2}
...

Question: {question}

Based on all agent summaries, answer the question.
ANSWER:
```

Free-form debate — initial round system instruction.

```
You are a question-answering assistant in a multi-agent reasoning team.
Answer based only on your provided documents.
Output your final answer as:
ANSWER: <answer>
```

Free-form debate — initial round user prompt.

```
Your private evidence:
{context}

Question: {question}

ANSWER:
```

Free-form debate — update round system instruction.

```
You are a question-answering assistant in a multi-agent reasoning team.
You can see your own evidence and what other agents answered in the
previous round. Consider whether any other agent's answer reveals
evidence you lacked.
Output your updated final answer as:
ANSWER: <answer>
```

Free-form debate — update round user prompt.

```
Your private evidence:
{context}

Other agents' answers (Round {round_num - 1}):
  Agent 1: {answer_1}
  Agent 2: {answer_2}
  ...

Question: {question}

Update your answer considering all available information.
ANSWER:
```

Confidence-gated — local answer + confidence system instruction.

```
You are a question-answering assistant.
Answer using ONLY the provided documents --- do not use background
knowledge. After your answer, rate your confidence from 0.0 to 1.0
that the answer is correct.
Use this exact format:
ANSWER: <answer>
CONFIDENCE: <0.0-1.0>
```

Confidence-gated — local answer + confidence user prompt.

```
Documents:
{context}

Question: {question}

Answer using only the documents above.
ANSWER: <answer>
CONFIDENCE: <0.0-1.0>
```

Confidence-gated / Disagreement-gated — aggregation system instruction.

```
You are a question-answering assistant. You have received evidence
from agents that could not answer locally. Use all provided
information to give the correct answer.
Output your final answer as:
ANSWER: <answer>
```

C Experiments and Additional Analyses

C.1 Experiment Setup

Benchmarks. We evaluate on two benchmarks. **Silo-Bench** comprises 52 tasks spanning two agent-count tiers (n=5 and n=10); results are reported as the combined average across both tiers. **MuSiQue** (distributed variant) comprises 100 instances: 50 two-hop, 35 three-hop, and 15 four-hop questions, with each supporting paragraph assigned to a separate agent.

Models and protocols. We evaluate ten LLMs (Table 11) across eleven communication protocols (three no-communication baselines, four natural-language protocols, three relay protocols, and Full Evidence Sharing as an oracle reference), all accessed via the OpenRouter API.

Relay protocol hyperparameters. The per-agent communication budget is k=3 evidence units for all relay experiments unless otherwise noted. Score-Ranked Relay uses equal weights for the three score components (relevance, uniqueness, local importance). Redundancy-Aware Relay applies a diversity penalty ρ=0.5 on pairwise cosine similarity between unit embeddings. All scoring calls use temperature 0.0; all answer-generation calls use temperature 0.7.

Evaluation. Accuracy is measured by token-level F1 against the gold answer string (Rajpurkar et al., 2016), normalized for capitalization and punctuation. Critical recall, omission rate, and aggregation failure metrics are computed per instance and averaged. Alignment judgements use Qwen-235B as the judge model. Communication tokens (CommTok) count the total tokens transmitted between agents; total token cost (TokCost) counts all LLM API tokens consumed across extraction, scoring, and aggregation steps.

Claim-to-evidence map. Table 12 maps each major empirical claim to its supporting table, the scope of evidence, and the primary limitation.

Table 12: Claim-to-evidence map. Each row is one empirical claim in the paper, with its supporting table and the scope of evidence. The map covers the four main findings: protocol comparison, failure-mode taxonomy, bottleneck structure, and local importance failure.

Claim	**Supporting table**	**Scope**
Full evidence sharing performs strongly for capable models; relay offers a performance–cost trade-off	Tab. 2	10 models $\times$ 2 benchmarks, 8–11 protocols
Silo-Bench relay shows incomplete measured critical coverage, with stage-wise extraction and selection losses	Tab. 3	9 models, 52 Silo-Bench tasks
MuSiQue errors persist despite complete lexical coverage for relay in 8 of 9 models; semantic sufficiency remains unverified	Tab. 3	9 models, 100 MuSiQue instances
Summary Exchange distorts or loses numeric evidence even when numbers appear in the summary	App. C.3.2	Qwen-235B, 390 (agent, gold-shard) pairs on Silo-Bench
Removing one annotated critical unit reduces accuracy; distractor and redundancy controls have model-dependent effects	Tab. 1	7 models, 52 Silo-Bench tasks
Local importance scores have moderate AUC (0.50–0.89 depending on model/score) but cannot reliably identify jointly-critical facts	Fig. 4	6 models, Silo-Bench only
Under tight budgets, low-score critical units are disproportionately dropped (FNR 77.5% vs. 30.8%)	App. C.3.1	Qwen-235B, 1,170 units, hypothetical $k{=}1$ scenario
Incomplete SB coverage and residual MQ errors recur across audited models, with exceptions	Tab. 3	9 models (8 with MQ CritRec = 1); both benchmarks
Score decomposition distinguishes uniqueness from relevance at tight budgets	App. C.3.1	Qwen-235B, Silo-Bench, $k{=}1$
Relay-to-full-sharing comparison improves task scores while exposing residual failures	Tab. 8	Qwen-235B, both benchmarks; retrospective comparison
Selected residual errors include reasoning and answer-matching discrepancies	App. C.4.4	61 selected MuSiQue answers; exploratory judge labels
Relay preserves numeric evidence more faithfully than summaries in the saved distortion audit	Tab. 39	Qwen-235B, Silo-Bench; unequal audit pair counts

Table 13: Full metrics — Qwen-235B (`qwen/qwen3-235b-a22b-2507`), Silo-Bench.

Protocol	Acc %↑	CommTok↓	TokCost↓	Utility↑	Recovery↑
No-communication bounds					
Single Local	3.85	0	1,868	0.038	0.000
Majority Vote	7.69	0	13,601	0.077	0.200
Best-Local Oracle	15.38	0	13,601	0.154	0.600
Natural-language communication					
Free-form Debate	15.38	13,364	56,732	−1.183	0.600
Summary Exchange	9.62	529	7,684	0.043	0.300
Confidence-Gated	7.69	352	7,517	0.042	0.200
Disagreement-Gated	15.38	918	16,716	0.062	0.600
Structured evidence relay					
Random Relay	17.31	698	8,127	0.103	0.700
Score-Ranked	19.23	698	18,036	0.123	0.800
Redundancy-Aware	19.23	698	18,155	0.123	0.800
Pooled-evidence reference					
Full Evidence Sharing	23.08	1,655	6,089	0.065	1.000

Table 14: Full metrics — Mistral-S (`mistralai/mistral-small-2603`), Silo-Bench.

Protocol	Acc %↑	CommTok↓	TokCost↓	Utility↑	Recovery↑
No-communication bounds					
Single Local	3.85	0	1,106	0.038	0.000
Majority Vote	7.69	0	7,894	0.077	0.286
Best-Local Oracle	9.62	0	7,894	0.096	0.429
Natural-language communication					
Free-form Debate	9.62	1,231	21,219	−0.027	0.429
Summary Exchange	13.46	594	6,571	0.075	0.714
Confidence-Gated	15.38	653	7,649	0.089	0.857
Disagreement-Gated	17.31	1,394	11,610	0.034	1.000
Structured evidence relay					
Random Relay	15.38	657	8,340	0.088	0.857
Score-Ranked	13.46	657	18,989	0.069	0.714
Redundancy-Aware	15.38	657	19,014	0.088	0.857
Pooled-evidence reference					
Full Evidence Sharing	17.31	1,655	4,524	0.008	1.000

C.2 Extended Results

C.2.1 Per-Model Full Results

Tables 13–18 extend the main Table 2 with communication volume, total model tokens, utility, and recovery. **TokCost** = mean total LLM API tokens consumed across all agents and all rounds (extraction, scoring, aggregation). **Utility** $= \text{Acc}/100 - 10^{-4} \cdot \text{CommTok}$, a scalar trade-off between accuracy and communication overhead. Task accuracy corresponds to the Silo-Bench scores in main Table 2. All tables report results on the Silo-Bench combined n=5+n=10 setup.

C.2.2 Multi-Model Evidence-Flow Audit

Table 3 in the main paper summarises the key audit metrics (CritRec and AggFail) for three representative models across both benchmarks. Tables 19–23 provide the full six-metric breakdown for the five original models, showing Silo-Bench and MuSiQue results side by side.

Table 15: Full metrics — GPT-OSS (`openai/gpt-oss-safeguard-20b`), Silo-Bench. Recovery > 1 indicates the protocol outperforms Full Evidence Sharing (which hurts this model).

Protocol	Acc %↑	CommTok↓	TokCost↓	Utility↑	Recovery↑
No-communication bounds					
Single Local	1.92	0	185	0.019	0.000
Majority Vote	5.77	0	1,658	0.058	2.003
Best-Local Oracle	7.69	0	1,658	0.077	3.003
Natural-language communication					
Free-form Debate	5.77	69	2,446	0.051	2.003
Summary Exchange	0.00	69	931	−0.007	−1.003
Confidence-Gated	3.85	1,580	819	−0.119	1.003
Disagreement-Gated	7.69	151	1,734	0.062	3.005
Structured evidence relay					
Random Relay	0.00	54	819	−0.005	−1.003
Score-Ranked	0.00	54	1,646	−0.005	−1.003
Redundancy-Aware	0.00	54	1,695	−0.005	−1.003
Pooled-evidence reference					
Full Evidence Sharing	3.85	1,655	179	−0.127	1.000

Table 16: Full metrics — Qwen-80B-T (`qwen/qwen3-next-80b-a3b-thinking`), Silo-Bench.

Protocol	Acc %↑	CommTok↓	TokCost↓	Utility↑	Recovery↑
No-communication bounds					
Single Local	0.00	0	5,686	0.000	0.000
Majority Vote	7.69	0	43,947	0.077	0.125
Best-Local Oracle	11.54	0	43,947	0.115	0.188
Natural-language communication					
Free-form Debate	9.62	576	125,754	0.039	0.156
Summary Exchange	19.23	374	30,511	0.155	0.313
Confidence-Gated	15.38	464	65,251	0.108	0.250
Disagreement-Gated	26.92	882	52,864	0.181	0.438
Structured evidence relay					
Random Relay	17.31	241	43,325	0.149	0.281
Score-Ranked	23.08	241	103,318	0.207	0.375
Redundancy-Aware	17.31	241	103,503	0.149	0.281
Pooled-evidence reference					
Full Evidence Sharing	61.54	1,655	16,084	0.450	1.000

Across all models with complete extraction data, the qualitative patterns hold: (i) Silo-Bench relay protocols achieve Crit. Rec. < 1 (ranging 0.71–0.93), indicating measured evidence loss across model families; (ii) MuSiQue relay protocols achieve Crit. Rec. $= 1$ for four of the five models, yet the reported AF proxy remains high (47–82%), showing that measured coverage alone does not guarantee correct answers. These legacy summaries should be interpreted alongside the binary-versus-mean-F1 distinction in Appendix C.4.4.

Ring-1T (`inclusionai/ring-2.6-1t`) had low extraction success rates (~38% for n=5; near zero for n=10), so its Silo-Bench relay audit is omitted; only Summary Exchange and Full Evidence Sharing rows are reported for that benchmark.

C.2.3 Stage-Wise Evidence Survival

Tables 6 (main paper) and 24 (below) show how gold critical evidence survives (or is lost) through successive pipeline stages for four models. We track four checkpoints:

Table 17: Full metrics — Ring-1T (`inclusionai/ring-2.6-1t`), Silo-Bench. TokCost†: n=10 API token logs were not collected for relay protocols and Full Evidence Sharing; TokCost for those rows reflects n=5 only.

Protocol	Acc %↑	CommTok↓	TokCost↓	Utility↑	Recovery↑
No-communication bounds					
Single Local	7.69	0	475	0.077	0.000
Majority Vote	5.77	0	2,960	0.058	−0.143
Best-Local Oracle	7.69	0	2,960	0.077	0.000
Natural-language communication					
Free-form Debate	5.77	23	2,975	0.055	−0.143
Summary Exchange	3.85	77	1,274	0.031	−0.286
Confidence-Gated	1.92	1,551	529†	−0.136	−0.429
Disagreement-Gated	7.69	359	3,039	0.041	0.000
Structured evidence relay					
Random Relay	5.77	66	2,009†	0.051	−0.143
Score-Ranked	5.77	66	2,569†	0.051	−0.143
Redundancy-Aware	3.85	66	2,462†	0.032	−0.286
Pooled-evidence reference					
Full Evidence Sharing	21.15	1,655	2,548†	0.046	1.000

Table 18: Full metrics — Qwen-3.6-Flash (`qwen/qwen3.6-flash`), Silo-Bench.

Protocol	Acc %↑	CommTok↓	TokCost↓	Utility↑	Recovery↑
No-communication bounds					
Single Local	7.69	0	7,482	0.077	0.000
Majority Vote	5.77	0	60,667	0.058	−0.033
Best-Local Oracle	13.46	0	60,667	0.135	0.100
Natural-language communication					
Free-form Debate	17.31	1,082	142,040	0.065	0.167
Summary Exchange	28.85	599	28,824	0.229	0.367
Confidence-Gated	9.62	84	40,226	0.088	0.033
Disagreement-Gated	59.62	1,472	75,107	0.449	0.900
Structured evidence relay					
Random Relay	61.54	760	35,019	0.539	0.933
Score-Ranked	59.62	760	83,965	0.520	0.900
Redundancy-Aware	61.54	760	84,921	0.539	0.933
Pooled-evidence reference					
Full Evidence Sharing	**65.38**	1,655	16,466	0.488	1.000

(1) **Sharding**: every critical unit resides in its agent's shard by construction; CritRec = 1.000.
(2) **Extraction**: CritRec over all units extracted by the shard agent (before top-k selection); captures evidence omission at the sharding/extraction boundary.
(3) **Transmission**: CritRec of the top-k units sent to the board; loss relative to extraction indicates selection omission.
(4) **Aggregation Failure (%)**: fraction of all instances with complete measured coverage and a non-perfect answer.

On Silo-Bench (Table 6), evidence omission begins at the extraction stage (CritRec = 0.80–0.93) and local importance failure adds a further 3–6 point drop. Candidate aggregation failures account for 27–71% of all instances. On MuSiQue (Table 24), measured extraction coverage is complete, with AF of 48–82%. This flags residual errors but does not establish lossless semantic transmission.

Table 19: Evidence-flow audit for **Qwen-235B** on both benchmarks. **CR**↑ = Critical Recall, **Om**↓ = Omission Rate, **Di**↓ = Distortion Rate, **Re**↓ = Redundancy, **UB**↑ = Useful Budget, **AF**↓ = Agg. Fail (%). Full metric definitions in Table 3.

	Silo-Bench						MuSiQue					
Protocol	**CR**↑	**Om**↓	**Di**↓	**Re**↓	**UB**↑	**AF**↓	**CR**↑	**Om**↓	**Di**↓	**Re**↓	**UB**↑	**AF**↓
Summary Exchange	0.669	0.331	—	0.328	0.379	53.8	1.000	0.000	—	0.221	0.776	79.0
Random Relay	0.844	0.156	0.074	0.167	0.255	67.3	1.000	0.000	0.030	0.151	0.319	72.0
Score-Ranked Relay	0.844	0.156	0.056	0.167	0.254	65.4	1.000	0.000	0.019	0.151	0.319	68.0
Redundancy-Aware Relay	0.844	0.156	0.060	0.167	0.254	65.4	1.000	0.000	0.023	0.151	0.319	72.0
Full Evidence Sharing	1.000	0.000	0.000	—	—	76.9	1.000	0.000	0.000	—	—	47.5

Table 20: Evidence-flow audit for **Mistral-S** on both benchmarks. Column abbreviations and metric definitions as in Table 19.

	Silo-Bench						MuSiQue					
Protocol	**CR**↑	**Om**↓	**Di**↓	**Re**↓	**UB**↑	**AF**↓	**CR**↑	**Om**↓	**Di**↓	**Re**↓	**UB**↑	**AF**↓
Summary Exchange	0.750	0.250	—	0.243	0.502	63.5	1.000	0.000	—	0.159	0.757	76.0
Random Relay	0.885	0.115	0.030	0.118	0.287	71.2	1.000	0.000	0.045	0.126	0.322	79.0
Score-Ranked Relay	0.885	0.115	0.030	0.118	0.286	69.2	1.000	0.000	0.042	0.126	0.326	82.0
Redundancy-Aware Relay	0.885	0.115	0.032	0.118	0.286	69.2	1.000	0.000	0.049	0.126	0.326	79.0
Full Evidence Sharing	1.000	0.000	0.000	—	—	82.7	1.000	0.000	0.000	—	—	47.7

Table 21: Evidence-flow audit for **Qwen-80B-T** on both benchmarks. Column abbreviations and metric definitions as in Table 19.

	Silo-Bench						MuSiQue					
Protocol	**CR**↑	**Om**↓	**Di**↓	**Re**↓	**UB**↑	**AF**↓	**CR**↑	**Om**↓	**Di**↓	**Re**↓	**UB**↑	**AF**↓
Summary Exchange	0.712	0.288	—	0.218	0.462	53.8	1.000	0.000	—	0.154	0.792	67.0
Random Relay	0.748	0.252	0.079	0.183	0.240	55.8	1.000	0.000	0.045	0.126	0.482	78.0
Score-Ranked Relay	0.748	0.252	0.051	0.183	0.245	50.0	1.000	0.000	0.038	0.126	0.480	75.0
Redundancy-Aware Relay	0.748	0.252	0.065	0.183	0.245	53.8	1.000	0.000	0.030	0.126	0.480	76.0
Full Evidence Sharing	1.000	0.000	0.000	—	—	38.5	1.000	0.000	0.000	—	—	47.7

Table 22: Evidence-flow audit for **Qwen-3.6-Flash** on both benchmarks. Column abbreviations and metric definitions as in Table 19.

	Silo-Bench						MuSiQue					
Protocol	**CR**↑	**Om**↓	**Di**↓	**Re**↓	**UB**↑	**AF**↓	**CR**↑	**Om**↓	**Di**↓	**Re**↓	**UB**↑	**AF**↓
Summary Exchange	0.725	0.275	—	0.287	0.477	50.0	1.000	0.000	—	0.145	0.874	65.0
Random Relay	0.844	0.156	0.072	0.139	0.273	28.8	1.000	0.000	0.030	0.144	0.360	53.0
Score-Ranked Relay	0.844	0.156	0.081	0.139	0.271	32.7	1.000	0.000	0.023	0.144	0.360	48.0
Redundancy-Aware Relay	0.844	0.156	0.063	0.139	0.271	26.9	1.000	0.000	0.023	0.144	0.360	51.0
Full Evidence Sharing	1.000	0.000	0.000	—	—	34.6	1.000	0.000	0.000	—	—	26.3

Qwen-3.6-Flash has lower AF (27–33% on Silo-Bench; 48–53% on MuSiQue); attributing this difference specifically to aggregation ability requires validating evidence sufficiency.

C.2.4 Critical-Evidence Bottleneck

Multi-model bottleneck on Silo-Bench. Table 1 in the main paper reports the critical-evidence bottleneck experiment for seven models on Silo-Bench. Following the same protocol, we construct

Table 23: Evidence-flow audit for **Ring-1T** on both benchmarks. Column abbreviations and metric definitions as in Table 19. Silo-Bench SE and FS rows use the n=10 split (26 tasks); relay rows (†) use the n=5 split (n=13 instances with successful extraction) because Ring-1T's structured extraction fails for all n=10 instances. Di is unavailable for relay rows (distortion audit not run for this model).

	Silo-Bench						MuSiQue					
Protocol	**CR↑**	**Om↓**	**Di↓**	**Re↓**	**UB↑**	**AF↓**	**CR↑**	**Om↓**	**Di↓**	**Re↓**	**UB↑**	**AF↓**
Summary Exchange	0.550	0.450	—	0.249	0.397	37.5	1.000	0.000	—	0.121	0.811	100.0
Random Relay†	0.692	0.308	—	0.103	0.210	75.0	0.967	0.033	0.052	0.132	0.499	96.7
Score-Ranked Relay†	0.692	0.308	—	0.103	0.210	75.0	0.967	0.033	0.052	0.132	0.483	96.7
Redundancy-Aware Relay†	0.692	0.308	—	0.103	0.210	87.5	0.967	0.033	0.052	0.132	0.483	96.7
Full Evidence Sharing	1.000	0.000	0.000	—	—	100.0	1.000	0.000	0.000	—	—	100.0

† Silo-Bench relay metrics from n=5 split only (n=13 instances); Ring-1T's scoring phase returns empty output, so Score-Ranked and Redundancy-Aware relay fall back to random selection, making their CR/Om/Red/UB identical to Random Relay.

Table 24: Stage-wise critical evidence survival on **MuSiQue** (100 instances). Extraction- and transmission-stage CritRec are both 1.000 for all models and protocols, so evidence omission is not the bottleneck; aggregation failure (48–82%) dominates.

Model	**Protocol**	**Raw**	**Extracted**	**Transmitted**	**AggFail (%)**
Qwen-235B	Random Relay	1.000	1.000	1.000	72.0
	Score-Ranked Relay	1.000	1.000	1.000	68.0
	Redundancy-Aware Relay	1.000	1.000	1.000	72.0
Mistral-S	Random Relay	1.000	1.000	1.000	79.0
	Score-Ranked Relay	1.000	1.000	1.000	82.0
	Redundancy-Aware Relay	1.000	1.000	1.000	79.0
Qwen-80B	Random Relay	1.000	1.000	1.000	78.0
	Score-Ranked Relay	1.000	1.000	1.000	75.0
	Redundancy-Aware Relay	1.000	1.000	1.000	76.0
Qwen-3.6-F	Random Relay	1.000	1.000	1.000	53.0
	Score-Ranked Relay	1.000	1.000	1.000	48.0
	Redundancy-Aware Relay	1.000	1.000	1.000	51.0

All relay protocols transmit all critical units on MuSiQue (CritRec = 1.000 at every stage), yet 48–82% of answers are still wrong. Aggregation failure, not evidence omission, is the dominant bottleneck on this benchmark. Qwen-3.6-Flash's lower AggFail (48–53%) reflects stronger aggregation ability rather than a qualitatively different failure mode.

an oracle evidence board containing each agent's full shard, one redundant copy, and one distractor unit, then systematically delete specific unit types and re-run the aggregation step. All deletions are performed from the oracle board (Full Evidence Sharing setting), isolating the aggregator's sensitivity to missing evidence from any relay-protocol effects.

Across all models, removing one or two critical units substantially reduces accuracy. This confirms that the causal relationship between critical evidence and final accuracy holds across model families. The drop is largest for Qwen-80B-T (a thinking model), which loses 67–79% relative accuracy when critical units are removed. For non-critical unit removal, Qwen-80B-T sees slightly *improved* accuracy when distractors are removed, suggesting thinking models are more sensitive to irrelevant context.

Among the additional models, Gemma-4-31B shows the sharpest drops (67% relative for one critical unit removed; 71% for two), comparable to the thinking-model pattern despite not using explicit chain-of-thought. GPT-4o-mini shows a clean dissociation: 38% relative drop for critical removal, zero change for non-critical removal. DeepSeek-R1-32B scores 0.0% in all conditions, consistent with its near-zero accuracy on relay protocols.

Table 25: Critical-evidence bottleneck on MuSiQue: accuracy (%) under controlled deletion from the oracle evidence board. Δ = absolute change from no-deletion baseline. All models show sharp accuracy decline when critical units are removed, confirming that aggregation failure is driven by missing evidence on multi-hop QA. Bootstrap 95% confidence intervals (10,000 resamples) do not overlap between any model's k=0 and k=1 conditions, confirming statistical significance of all reported drops.

	Critical units removed			**Non-critical removed**	
Model	k=0	k=1	k=2	−Distractor	−Redundant
Qwen-235B	48.8	28.9 (−19.9)	8.5 (−40.3)	52.1	49.0
Mistral-S	38.6	21.4 (−17.2)	3.4 (−35.2)	37.4	35.4
Qwen-80B-T	52.2	16.7 (−35.6)	2.3 (−49.9)	49.5	53.2
Qwen-3.6-F	68.4	31.4 (−37.0)	10.4 (−58.0)	72.1	73.2
Gemini-FL	32.2	20.0 (−12.2)	5.4 (−26.8)	38.5	30.4
GPT-4o-m	38.8	22.7 (−16.1)	5.0 (−33.8)	41.3	39.5
Gemma-4-31B	48.9	14.0 (−34.8)	1.8 (−47.1)	56.0	55.1

Table 26: Multi-model hop-count error summaries on MuSiQue (100 instances), for Full Evidence Sharing and Random Relay. For Full Evidence Sharing, the saved entries are $100 - \mathrm{F1}$, not binary aggregation-failure rates; they should not be compared directly with relay AF. For relay, AF is the percentage of all instances with a non-perfect answer and complete measured CritRec at transmission. CritRec = 1.000 across all relay protocols for Qwen-235B, Mistral-S, and Qwen-80B-T; Qwen-3.6-Flash also achieves CritRec = 1.000 on all relay protocols. Qwen-3.6-Flash has higher full-sharing mean F1 (73.7 vs. ≈52 for the other models).

Protocol	**Model**	**2-hop**	**3-hop**	**4-hop**	**All hops**
Full Sharing ($100 - \mathrm{F1}$)	Qwen-235B	39.7	50.6	66.3	47.5
	Mistral-S	41.4	46.6	71.3	47.7
	Qwen-80B	41.7	48.7	65.3	47.7
	Qwen-3.6-F	29.5	28.4	11.0	26.3
Random Relay	Qwen-235B	62.0	82.9	80.0	72.0
	Mistral-S	72.0	91.4	73.3	79.0
	Qwen-80B	66.0	85.7	100.0	78.0
	Qwen-3.6-F	54.0	54.3	46.7	53.0

Critical-evidence bottleneck on MuSiQue. Table 25 extends the bottleneck experiment to MuSiQue using the same deletion protocol. The bottleneck pattern is even sharper: removing one critical paragraph reduces accuracy by 12–37 absolute points (25–68% relative); removing two collapses most models below 11%. This large sensitivity is consistent with the causal structure of multi-hop reasoning: removing any single gold paragraph breaks one link in the reasoning chain. Non-critical deletions show no systematic negative effect; several models improve when distractors are removed.

C.2.5 Hop-Count Stratification on MuSiQue

Table 5 (main paper) breaks down MuSiQue results by hop count for Qwen-235B. Table 26 below extends this to all four models with complete data (AggFail summary only).

Two patterns emerge from the Qwen-235B detailed table. First, mean answer F1 falls with hop count under Full Evidence Sharing, from 60.3 at two hops to 33.7 at four hops. Second, relay AF is higher in the three- and four-hop groups than in the two-hop group, although it need not rise monotonically. Full-sharing entries in the supplemental AF table below are $100 - \mathrm{F1}$, not binary AF, and must not be compared directly with relay AF. CritRec = 1.000 holds under the permissive lexical matcher across these relay strata; this does not rule out missing or altered meaning.

Table 27: Budget k-sweep on MuSiQue across five models. Each relay variant transmits up to k units per agent; extraction caps at $k_{\max}$=3. Score-Ranked shows a consistent advantage over Random at k=1 (average gap +12.1 points), confirming the budget-sensitivity benefit of local ranking on multi-hop QA. The gap narrows substantially at k=3, consistent with the joint-criticality explanation. Acc (%); Q235B = Qwen-235B; Mis-S = Mistral-S; Q3.6F = Qwen-3.6-Flash; Gem-31B = Gemma-4-31B; Q80B-T = Qwen-80B-T.

	k=1					k=2					k=3				
Protocol	Q235B	Mis-S	Q3.6F	Gem-31B	Q80B-T	Q235B	Mis-S	Q3.6F	Gem-31B	Q80B-T	Q235B	Mis-S	Q3.6F	Gem-31B	Q80B-T
Random Relay	23.8	8.4	30.7	8.8	17.7	33.7	25.3	52.5	22.7	24.4	42.7	30.5	62.8	37.0	29.3
Score-Ranked Relay	35.0	16.6	51.0	29.8	16.8	43.5	31.7	61.4	33.1	29.5	43.8	29.8	65.1	37.0	34.5
Gap (Score−Rand)	*+11.2*	*+8.2*	*+20.3*	*+21.0*	*−0.8*	*+9.9*	*+6.4*	*+8.9*	*+10.4*	*+5.1*	*+1.1*	*−0.7*	*+2.3*	*0.0*	*+5.2*

Table 28: Evidence-flow audit for **Gemini-FL** on both benchmarks. Column abbreviations and metric definitions as in Table 19. Distortion metrics omitted for this model (require separate LLM judge calls).

	Silo-Bench						**MuSiQue**					
Protocol	**CR↑**	**Om↓**	**Di↓**	**Re↓**	**UB↑**	**AF↓**	**CR↑**	**Om↓**	**Di↓**	**Re↓**	**UB↑**	**AF↓**
Summary Exchange	0.687	0.313	—	0.405	0.369	51.9	1.000	0.000	—	0.190	0.745	92.0
Random Relay	0.935	0.065	—	0.155	0.312	69.2	1.000	0.000	—	0.114	0.331	78.0
Score-Ranked Relay	0.935	0.065	—	0.155	0.312	71.2	1.000	0.000	—	0.114	0.330	77.0
Redundancy-Aware Relay	0.935	0.065	—	0.155	0.312	73.1	1.000	0.000	—	0.114	0.330	77.0
Full Evidence Sharing	1.000	0.000	—	—	—	80.8	1.000	0.000	—	—	—	65.0

The multi-model comparison (Table 26) shows that the full-sharing F1 complement rises with hop count for Qwen-235B, Mistral-S, and Qwen-80B-T. Qwen-3.6-Flash is an exception: its F1 complement falls with hop count in this sample, so the trend is model-dependent.

C.2.6 Budget k-Sweep on MuSiQue

The budget k-sweep on Silo-Bench is reported as Table 4 in the main paper. Table 27 below extends the analysis to MuSiQue, covering the same five models at k=1, 2, 3. Score-Ranked Relay outperforms Random Relay by an average of 12.1 points at k=1 across five models (vs. 15.7 on Silo-Bench), with the largest gains for Qwen-3.6-Flash (+20.3) and Gemma-4-31B (+21.0). This confirms that the benefit of local importance ranking is not benchmark-specific. At k=3, the gap narrows to 1–5 points for most models, consistent with the joint-criticality explanation: at tighter budgets, the top-ranked unit is more likely to be locally decisive; at looser budgets, marginal units are jointly-critical and local scoring cannot reliably identify them.

C.2.7 Additional Model Robustness Evaluation

This section reports evidence-flow audit metrics for four additional models: Gemini-2.5-Flash-Lite, GPT-4o-mini, DeepSeek-R1-distill-Qwen-32B, and Gemma-4-31B-it. All models were evaluated on both benchmarks using the same protocols, prompts, and hyperparameters as the main paper. Protocol comparison accuracy for all four models is included in Table 2. Combined Silo-Bench and MuSiQue audit results are in Tables 28–31.

Hop-count stratification for additional models. Table 32 extends the hop-count stratification (Appendix C.2.5) to the four additional models.

C.3 Additional Analyses

C.3.1 Local Scoring and Joint-Criticality Analysis

Figure 4 (main paper) summarises the local importance failure analysis across models; Table 33 below presents a budget-sensitivity analysis linking local importance failure to evidence omission.

Table 29: Evidence-flow audit for **GPT-4o-mini** on both benchmarks. Column abbreviations and metric definitions as in Table 19. Distortion metrics omitted (require separate LLM judge calls).

	Silo-Bench						MuSiQue					
Protocol	**CR↑**	**Om↓**	**Di↓**	**Re↓**	**UB↑**	**AF↓**	**CR↑**	**Om↓**	**Di↓**	**Re↓**	**UB↑**	**AF↓**
Summary Exchange	0.700	0.300	—	0.337	0.444	61.5	1.000	0.000	—	0.178	0.880	73.0
Random Relay	0.883	0.117	—	0.151	0.280	73.1	1.000	0.000	—	0.115	0.330	70.0
Score-Ranked Relay	0.883	0.117	—	0.151	0.279	71.2	1.000	0.000	—	0.115	0.333	71.0
Redundancy-Aware Relay	0.883	0.117	—	0.151	0.279	71.2	1.000	0.000	—	0.115	0.333	74.0
Full Evidence Sharing	1.000	0.000	—	—	—	80.8	1.000	0.000	—	—	—	67.0

Table 30: Evidence-flow audit for **DeepSeek-32B** on both benchmarks. Column abbreviations and metric definitions as in Table 19. Distortion metrics omitted (require separate LLM judge calls). Low CR values reflect poor extraction success on both benchmarks; this model is a degenerate case.

	Silo-Bench						MuSiQue					
Protocol	**CR↑**	**Om↓**	**Di↓**	**Re↓**	**UB↑**	**AF↓**	**CR↑**	**Om↓**	**Di↓**	**Re↓**	**UB↑**	**AF↓**
Summary Exchange	0.471	0.529	—	0.000	0.362	22.0	0.269	0.731	—	0.000	0.302	8.5
Random Relay	0.595	0.405	—	0.000	0.624	47.3	0.426	0.574	—	0.000	0.424	15.3
Score-Ranked Relay	0.595	0.405	—	0.000	0.624	47.3	0.426	0.574	—	0.000	0.424	15.3
Redundancy-Aware Relay	0.595	0.405	—	0.000	0.624	47.3	0.426	0.574	—	0.000	0.424	15.3
Full Evidence Sharing	1.000	0.000	—	—	—	88.5	1.000	0.000	—	—	—	100.0

Table 31: Evidence-flow audit for **Gemma-4-31B** on both benchmarks. Column abbreviations and metric definitions as in Table 19. Distortion metrics omitted (require separate LLM judge calls).

	Silo-Bench						MuSiQue					
Protocol	**CR↑**	**Om↓**	**Di↓**	**Re↓**	**UB↑**	**AF↓**	**CR↑**	**Om↓**	**Di↓**	**Re↓**	**UB↑**	**AF↓**
Summary Exchange	0.850	0.150	—	0.279	0.727	32.7	1.000	0.000	—	0.139	0.803	72.0
Random Relay	0.883	0.117	—	0.200	0.285	50.0	1.000	0.000	—	0.135	0.438	69.0
Score-Ranked Relay	0.883	0.117	—	0.200	0.285	53.8	1.000	0.000	—	0.135	0.436	74.0
Redundancy-Aware Relay	0.883	0.117	—	0.200	0.285	42.3	1.000	0.000	—	0.135	0.436	73.0
Full Evidence Sharing	1.000	0.000	—	—	—	44.2	1.000	0.000	—	—	—	58.0

Table 32: AggFail (%) by hop count on MuSiQue (100 instances) for four additional models (Full Evidence Sharing and Random Relay).

Protocol	**Model**	**2-hop**	**3-hop**	**4-hop**	**Pooled**
Full Evidence Sharing	Gemini-2.5-Flash-Lite	64.0	62.9	73.3	65.0
	GPT-4o-m	62.0	65.7	86.7	67.0
	DeepSeek-R1-32B	100.0	100.0	100.0	100.0
	Gemma-4-31B	50.0	65.7	66.7	58.0
Random Relay	Gemini-2.5-Flash-Lite	70.0	82.9	93.3	78.0
	GPT-4o-m	62.0	77.1	80.0	70.0
	DeepSeek-R1-32B	100.0	100.0	100.0	100.0
	Gemma-4-31B	70.0	65.7	73.3	69.0

Setup. We use 1,170 extracted unit records from 52 Silo-Bench instances (Qwen-235B, combined n=5+n=10). Each (instance, agent) group contains exactly $k_{\max}$=3 extracted units; the current protocol transmits all three (k=3). We simulate tighter budgets by selecting the top-k units per agent by `overall` score (the same score used by Score-Ranked Relay) and compute: (1) Critical Recall

Table 33: Joint-criticality false negative analysis (Silo-Bench, Qwen-235B, 1,170 units, n=5+n=10). **CritRec**: fraction of gold critical units retained at budget k. **FNR**: fraction of critical units dropped. **TP score / FN score**: mean `overall` local importance score for retained / dropped critical units. **Score gap**: TP score − FN score. At k=3 (current protocol) all units are transmitted; k=1 and k=2 are hypothetical tighter budgets.

Budget k	**CritRec**↑	**FNR**↓	**TP** n	**FN** n	**TP score**↑	**FN score**↑
k=1	0.661	0.391	137	88	0.935	0.880
k=2	0.899	0.107	201	24	0.923	0.835
k=3	1.000	0.000	225	0	0.913	—

Table 34: Multi-model local importance analysis on Silo-Bench (n=5+n=10, 52 tasks). **AUC**: area under the ROC curve for predicting gold criticality from each score. **Spearman**: rank correlation between score and gold label. **P@3 / R@3**: mean precision and recall at top-3 across instances. Gold critical rate: 19.2% (Qwen-235B, 1,170 units), 17.3% (Mistral-S, 1,167 units), 12.5% (Qwen-80B-T, 1,058 units), 17.6% (Qwen-3.6-F, 1,168 units).

	Qwen-235B				**Mistral-S**				**Qwen-80B**				**Qwen-3.6-F**			
Score	AUC	Spear.	P@3	R@3	AUC	Spear.	P@3	R@3	AUC	Spear.	P@3	R@3	AUC	Spear.	P@3	R@3
Relevance	0.501	+0.020	0.186	0.074	0.518	+0.080	0.179	0.080	0.570	+0.094	0.154	0.078	0.689	+0.297	0.308	0.132
Uniqueness	0.761	+0.362	0.436	0.179	0.795	+0.408	0.391	0.173	0.642	+0.178	0.199	0.102	0.881	+0.542	0.494	0.215
Importance	0.696	+0.277	0.314	0.129	0.729	+0.326	0.295	0.129	0.629	+0.162	0.167	0.086	0.871	+0.500	0.436	0.186
Overall	0.744	+0.334	0.391	0.164	0.746	+0.328	0.365	0.161	0.676	+0.210	0.199	0.101	0.917	+0.552	0.494	0.217

Qwen-3.6-F achieves the highest Overall AUC (0.917) and Spearman ρ (+0.552) of the original models; unusually, its Relevance AUC (0.689) is well above random, unlike other models where Relevance ≈ 0.5. The Uniqueness > Overall > Importance > Relevance ranking holds for Qwen-235B, Mistral-S, and Qwen-3.6-F, while Qwen-80B-T shows uniformly lower AUC values.

(CritRec): fraction of gold critical units transmitted; (2) False Negative Rate (FNR): fraction of gold critical units dropped; (3) Mean score of true-positive (TP) critical units: average score of selected critical units; (4) Mean score of false-negative (FN) critical units: average score of dropped critical units.

False negatives are disproportionately low-score critical units. At k=1, FNR = 39.1%: nearly two in five gold critical units would be dropped. The score gap between retained and dropped critical units is $0.935 - 0.880 = 0.055$—the local scorer does assign higher scores to critical units on average, but the gap is narrow.

Splitting critical units by score relative to the overall-score median ($\theta = 0.887$): among the 40 locally-low-score critical units (mean score 0.816), 77.5% become false negatives at k=1; among the 185 locally-high-score critical units (mean score 0.935), only 30.8% become false negatives. These are exactly the jointly-critical facts that a local scorer cannot identify.

Score decomposition. Table 7 shows the discriminative power of each score component at k=1, extending the scoring analysis in Section 4.3. *Uniqueness* has the largest TP–FN gap (0.270): syntactically distinctive values are both easy to identify locally and more likely to be truly critical. *Relevance* has zero gap (0.000) because almost every extracted unit is relevant to the question—it provides no discriminative signal. The combined overall score has a gap of only 0.055, meaning that even the best composite score leaves many jointly-critical facts below the selection threshold under tight budgets.

Multi-model local importance analysis. Table 34 extends the analysis to Mistral-S and Qwen-80B-T.

Table 35 further extends the analysis to Gemini-2.5-Flash-Lite, GPT-4o-mini, DeepSeek-R1-distill-Qwen-32B, and Gemma-4-31B-it.

Table 35: Local importance analysis on Silo-Bench (n=5+n=10, 52 tasks) for four additional models. Metrics as in Table 34. **AUC**: area under the ROC curve for predicting gold criticality. **Spearman**: rank correlation between score and gold label. **P@3 / R@3**: mean precision and recall at top-3. Gold critical rate: 17.2% (Gemini-Flash-Lite, 1,154 units), 18.0% (GPT-4o-mini, 1,169 units), 29.7% (DeepSeek-R1-32B, 192 units), 17.1% (Gemma-4-31B, 1,164 units). DeepSeek-R1-32B outputs a constant score for all units (AUC = 0.500; Spearman undefined, marked —); GPT-4o-mini's *Overall* score similarly collapses to near-random (AUC = 0.500).

	Gemini-Flash-Lite				GPT-4o-m				DeepSeek-R1-32B				Gemma-4-31B			
Score	AUC	Spear.	P@3	R@3	AUC	Spear.	P@3	R@3	AUC	Spear.	P@3	R@3	AUC	Spear.	P@3	R@3
Relevance	0.505	+0.028	0.167	0.075	0.502	+0.024	0.179	0.076	0.500	—	0.297	0.354	0.519	+0.076	0.179	0.080
Uniqueness	0.790	+0.403	0.333	0.151	0.541	+0.070	0.250	0.108	0.500	—	0.297	0.354	0.855	+0.517	0.468	0.207
Importance	0.728	+0.307	0.256	0.116	0.544	+0.131	0.179	0.076	0.500	—	0.297	0.354	0.791	+0.406	0.378	0.166
Overall	0.757	+0.342	0.244	0.106	0.500	−0.000	0.212	0.093	0.500	—	0.297	0.354	0.885	+0.513	0.468	0.207

Gemma-4-31B achieves the highest Uniqueness AUC (0.855) and Overall AUC (0.885) of any model tested, suggesting stronger local scoring capability. GPT-4o-mini's low Uniqueness AUC (0.541) and collapsed Overall score indicate poor local discriminability despite reasonable relay accuracy. DeepSeek-R1-32B produces constant scores across all units, making local importance scoring entirely uninformative for this model.

Local importance scoring on MuSiQue. Table 36 extends the analysis to MuSiQue. A sharp contrast with Silo-Bench emerges: local importance and overall composite scores are constant across units for all five models (AUC = 0.50; undefined Spearman), while uniqueness and relevance retain only marginal signal (AUC ≈ 0.49–0.64). The collapse is structural: on MuSiQue, each agent holds exactly one gold supporting paragraph, so there is no within-shard variation on which to base a local importance ranking. Compare to Silo-Bench, where uniqueness achieves AUC ≈ 0.64–0.88 across models (Table 34): the contrast confirms that local importance scoring is only informative when agents hold multi-unit shards with meaningful within-shard heterogeneity. At the current budget (k=3), complete measured relay coverage leaves little room for these scores to alter the coverage proxy. This does not establish that semantic evidence selection is solved; tighter budgets and validated fact-level coverage are needed to assess that claim.

Table 36: Local importance scoring quality on MuSiQue (100 instances, five models). **AUC**: area under the ROC curve for predicting gold criticality from each score component. † = constant input (all units receive identical scores from the model; Spearman undefined). Gold criticality is approximated via Jaccard matching of extracted units to gold supporting paragraphs. For reference, Silo-Bench uniqueness AUC values are: Qwen-235B = 0.761, Mistral-S = 0.795, Qwen-3.6-F = 0.881, Qwen-80B-T = 0.642 (Table 34).

Score	**Qwen-235B**	**Mistral-S**	**Qwen-3.6-F**	**Qwen-80B-T**	**Gemma-4-31B**
Relevance	0.503	0.532	0.635	0.569	0.591
Uniqueness	0.547	0.547	0.604	0.531	0.487
Local Importance	0.500†	0.500†	0.500†	0.500†	0.500†
Overall	0.500†	0.500†	0.500†	0.500†	0.500†

On MuSiQue, each agent holds exactly one gold supporting paragraph, providing no within-shard comparison basis. Local Importance and Overall collapse to constant outputs for all models. Compare to Silo-Bench Uniqueness AUC ≈ 0.76–0.88: the contrast confirms that local scoring is only informative when agents hold multi-unit shards requiring within-shard prioritisation.

Uniqueness-only ablation. To isolate the contribution of uniqueness to score-ranked relay, we ablate the composite score against a uniqueness-only selection strategy on Silo-Bench. Table 37 reports accuracy at k=1, 2, 3 for Qwen-235B and Qwen-3.6-F. This is a separate aggregation rerun on the same 52 tasks using saved extraction and scoring outputs, not a re-tabulation of the runs in Table 4. The scripts use different random-subset seeds and answer-normalization rules, and the new aggregation calls may vary; hence even their Random and Score-Ranked accuracy cells need not match across tables. Interpret each table's within-run comparisons, not cross-table differences, and note that repeated-run uncertainty is unavailable. Two findings emerge. First, *uniqueness-only*

Table 37: Uniqueness-only vs. composite score relay on Silo-Bench (combined n=5+n=10, 52 tasks). **Uniqueness Only** ranks units by uniqueness score alone; **Score Ranked** uses the composite overall score. At k=1, uniqueness-only outperforms the composite score for both models in this separate aggregation rerun. At k=3, all protocols transmit the same units; differences do not isolate selection and may reflect ordering or completion variability.

	Qwen-235B			**Qwen-3.6-F**		
Selection Strategy	k=1	k=2	k=3	k=1	k=2	k=3
Random Relay	11.5	17.3	19.2	11.5	15.4	48.1
Uniqueness Only	**17.3**	17.3	21.2	**38.5**	**40.4**	42.3
Score Ranked	13.5	**21.2**	19.2	19.2	38.5	**44.2**

Table 38: Distortion audit for Summary Exchange on Silo-Bench (Qwen-235B, 52 instances, n=5+n=10). Each row is one (agent summary, gold shard) pair. **Coverage**: fraction of pairs where at least one gold shard number appears in the summary. **Strict dist.**: fraction of *covered* pairs labelled DISTORTED (value wrong). **Broad degrad.**: fraction of covered pairs labelled PARTIAL or DISTORTED (value missing or wrong). **Exact/Para**: fraction of covered pairs with fully faithful transmission.

Split	N **pairs**	**Coverage↑**	**Strict dist.↓**	**Broad degrad.↓**	**Exact/Para↑**
n=5	130	0.385	10.0%	62.0%	38.0%
n=10	260	0.377	8.2%	63.3%	36.7%
Combined	390	0.379	8.8%	62.8%	37.2%

Label distribution (combined): NO_MATCH = 242 (62.1%), EXACT = 36 (9.2%), PARAPHRASE = 19 (4.9%), PARTIAL = 80 (20.5%), DISTORTED = 13 (3.3%).

ranking outperforms the composite score at k=1: 17.3% vs. 13.5% for Qwen-235B and 38.5% vs. 19.2% for Qwen-3.6-F, supporting its value as a local selection signal in this rerun. Second, *the composite score offers no consistent advantage*: it ties uniqueness-only at k=2 for Qwen-235B and is slightly lower for Qwen-3.6-F at k=2. This suggests that adding relevance (AUC $\approx$ 0.50) and local importance (AUC $\approx$ 0.70) to uniqueness does not provide additional discriminative power at tight budgets. At k=3, all three protocols transmit the same set of units per agent (since $k = k_{\max} = 3$); observed differences at this budget reflect prompt-ordering noise rather than selection quality.

C.3.2 EVIDENCE FIDELITY: AUDIT DETAILS AND MODEL COMPARISONS

Summary Exchange (Silo-Bench). Table 38 presents the evidence distortion audit for Summary Exchange (Qwen-235B, Silo-Bench). For each (agent summary, gold shard) pair (390 total across both n settings), we first check numeric coverage. For pairs with at least one numeric overlap (148 pairs), we call Qwen-235B as a judge to classify the relationship as EXACT, PARAPHRASE, PARTIAL, DISTORTED, or NO_MATCH.

Three findings emerge. First, most summaries omit numeric evidence entirely: 62.1% of pairs have no numeric overlap, explaining the low Crit. Rec. = 0.669 for Summary Exchange. Second, transmission is rarely exact: among the 37.9% with at least one gold value transmitted, only 37.2% are fully faithful (EXACT or PARAPHRASE). Third, the dominant failure is omission, not distortion: strict distortion affects 8.8% of covered pairs, while non-transmission affects 62.1% of all pairs.

Relay protocols vs. Summary Exchange (Silo-Bench). Table 39 extends the distortion audit to all three relay protocols at k=3, covering 315 (agent, gold shard) pairs per relay protocol, compared with 390 summary pairs.

Table 39: Evidence fidelity on Silo-Bench (Qwen-235B). Coverage: numeric overlap; Dist.: distorted among covered pairs; Faithful: exact/paraphrase among covered pairs. Different pair counts reflect the saved audit samples.

Protocol	Pairs	Coverage	Dist.	Faithful
Random Relay	315	97.1%	0.7%	75.2%
Score-Ranked Relay	315	97.5%	0.3%	72.3%
Redundancy-Aware Relay	315	97.5%	0.0%	73.3%
Summary Exchange	390	37.9%	8.8%	37.2%

Three findings emerge. First, relay dramatically outperforms Summary Exchange in coverage: 97.1–97.5% numeric coverage vs. 37.9%. Second, relay transmits evidence with high fidelity: 72–75% of covered pairs are rated EXACT or PARAPHRASE vs. 37.2% for Summary Exchange. Third, strict distortion is near-zero for relay: only 0.3–0.7% of covered pairs vs. 8.8% for Summary Exchange. These results support better numeric preservation by relay in the audited samples. Partial matches remain common, so they do not establish complete semantic coverage or a dominant aggregation failure mode.

For Random Relay, the full label distribution is 197 EXACT, 33 PARAPHRASE, 74 PARTIAL, two DISTORTED, and nine NO_MATCH pairs. Broad degradation (partial or distorted among covered pairs) is 24.8%, 27.7%, and 26.7% for random, score-ranked, and redundancy-aware relay, respectively, versus 62.8% for summaries. Numeric coverage differs from the lexical CritRec matcher, and the unequal pair counts mean these summaries are not a matched-pair causal comparison.

Multi-model relay distortion. Table 40 extends the score-ranked relay distortion audit to four additional models. Three of the four show the same pattern as Qwen-235B: high numeric coverage (89.5–96.9%) and low strict distortion (0.8–4.6%), confirming that the relay evidence-fidelity advantage is not model-specific. The Exact+Para rate of 77–84% for Gemini-FL, GPT-4o-mini, and Gemma-4-31B is slightly higher than Qwen-235B (72.3%), likely because these models extract more concise atomic units. DeepSeek-R1-32B is an outlier—coverage of only 35.1% indicates this model extracts non-numeric or poorly structured units, consistent with its near-zero accuracy on Silo-Bench tasks (§C.2.7).

Table 40: Distortion audit for Score-Ranked Evidence Relay across five models on Silo-Bench (n=5+n=10). Metrics as in Table 39. Qwen-235B from Table 39 (score-ranked relay row) is included for reference. DeepSeek-R1-32B shows low coverage because it extracts non-numeric units, consistent with its near-zero task accuracy.

Model	**Coverage↑**	**Strict dist.↓**	**Broad degrad.↓**	**Exact/Para↑**
Qwen-235B	97.5%	0.3%	27.7%	72.3%
Gemini-FL	96.9%	1.3%	15.6%	84.4%
GPT-4o-mini	89.5%	4.6%	22.6%	77.4%
Gemma-4-31B	91.3%	0.8%	19.4%	80.6%
DeepSeek-R1-32B	35.1%	10.6%	43.9%	56.1%
SE reference (Qwen-235B)	*37.9%*	*8.8%*	*62.8%*	*37.2%*

Table 41: Sensitivity of measured critical recall to the coverage threshold (Qwen-235B; SB: 26 five-agent instances; MQ: 100 instances). Coverage requires the maximum numeric-overlap / token-Jaccard score to be strictly greater than τ.

τ	SB CritRec	MQ CritRec	Recall ordering
0.0	0.885	1.000	Tied across relay variants
0.1	0.885	0.933	Tied across relay variants
0.2	0.785	0.687	Tied across relay variants
0.3	0.762	0.506	Tied across relay variants
0.5	0.685	0.267	Tied across relay variants

The three reconstructed relay variants have identical CritRec at every threshold: all select up to three units from the same extraction pool capped at three. Thus the tie is structural, rather than evidence of robust ranking among different selected sets.

C.4 Measurement and Interaction Controls

These analyses supplement the main audit with sensitivity checks and small controlled experiments. We distinguish experiments on the original Qwen-235B evaluation set from controls using other models or smaller samples. The source results and the table-generation procedure are documented alongside the paper.

C.4.1 Coverage-Threshold Sensitivity

We reconstruct transmitted evidence from cached Qwen-235B extraction and scoring outputs for 26 five-agent Silo-Bench instances and 100 MuSiQue instances. For a gold unit g and transmitted unit t, the coverage score is the maximum of token-set Jaccard overlap and the fraction of numeric strings in g that also occur in t; the numeric component is zero when g contains no numeric strings. A gold unit is covered when its best score over transmitted units is strictly greater than τ. The default is $\tau = 0$, as specified in Appendix B.6; this is a permissive lexical proxy, not a test of semantic entailment.

Table 41 shows that increasing τ from 0 to 0.5 reduces SB CritRec from 0.885 to 0.685 and MQ CritRec from 1.000 to 0.267. At $\tau = 0.1$, MQ still has higher measured recall (0.933 versus 0.885); from $\tau = 0.2$ onward, that ordering reverses. Consequently, the absolute recall values and the cross-benchmark contrast are not invariant to threshold choice. Comparing short atomic units against entire paragraphs can penalize faithful concise statements at stricter thresholds, while permissive overlap can also count incidental shared words or numbers as coverage. The existing deletion studies provide a complementary intervention-based check on the relevance of the gold supports, but do not validate every lexical alignment.

All three relay variants have identical recall in this reconstruction because the extraction pool and transmission budget are both capped at three units per agent. The variants can reorder those units but cannot select a different complete set. The unchanged recall ordering is therefore a tie induced by this setup, not a test of ranking stability under selective compression.

Table 42 reports the corresponding changes in AF, using the percentage of all evaluated instances rather than conditioning the denominator on complete coverage. For score-ranked relay, SB AF falls from 69.2% to 42.3% and MQ AF from 68.0% to 3.0% as the threshold increases. The predictions do not change: fewer wrong answers qualify as complete-coverage cases. This illustrates why a lower measured AF under a different matcher is not evidence of better aggregation. The SB values here concern only the five-agent tier and should not be confused with the combined 52-instance results in the main paper.

C.4.2 Zero-Evidence Controls

We give the aggregator the question and an empty evidence board while retaining the instruction to answer using only supplied evidence. Table 43 reports the recorded model identities and separates attempted instances from returned responses. The SB control uses DeepSeek-V4-Flash on 26 instances from an earlier ablation run; the MQ control uses Qwen3-8B on 100 attempted instances. These are auxiliary controls, not additional matched conditions for the ten-model protocol comparison.

Table 42: Aggregation-failure sensitivity on the same cached instances as Table 41. Each cell is the percentage of all instances with CritRec equal to 1 and a non-perfect answer.

Data	τ	Random relay	Score-ranked relay	Redundancy-aware relay
SB	0.0	73.1	69.2	69.2
SB	0.1	73.1	69.2	69.2
SB	0.2	53.8	50.0	50.0
SB	0.3	50.0	46.2	46.2
SB	0.5	46.2	42.3	42.3
MQ	0.0	72.0	68.0	72.0
MQ	0.1	58.0	56.0	59.0
MQ	0.2	24.0	23.0	26.0
MQ	0.3	16.0	14.0	16.0
MQ	0.5	3.0	3.0	3.0

The answers are held fixed. A lower AF at a stricter threshold means fewer instances meet the complete-coverage condition, not improved reasoning.

Table 43: Zero-evidence controls using the models recorded in the saved runs. Exact success is computed over nonempty responses with nonzero recorded token usage.

Data	Model	Attempted	Returned	Correct	Acc (%)
SB	DeepSeek-V4-Flash	26	26	0/26	0.0
MQ	Qwen3-8B	100	98	0/98	0.0

MQ has two empty responses with zero recorded token usage, excluded from the success denominator. Its 98 returned answers all abstain. The prompt explicitly restricts the model to supplied evidence, so this measures evidence-constrained answering with an empty board, not unrestricted parametric recall.

No returned answer is exactly correct in either control. On MuSiQue, all 98 returned predictions are "Insufficient information."; the remaining two entries contain an empty prediction and zero recorded token usage. The earlier attempted Qwen-235B MQ run contains only empty, zero-token entries and is excluded from model-performance evidence. The observed zeros establish failure to answer in this evidence-constrained setup. They do not establish that the models lack parametric knowledge of the answers: the instruction explicitly discourages using outside knowledge, and abstention is compatible with following that instruction. A direct test of parametric recall would require a separate question-only prompt that permits answering from prior knowledge, ideally on the same models and instances as the evidence-present comparison.

C.4.3 Fixed-Evidence-Budget Multi-Round Pilot

The pilot evaluates Qwen3-8B on five MuSiQue instances, holding the total transmission budget at three evidence units per agent. The one-round schedule sends three units; the two-round schedule sends one and then two; and the three-round schedule sends one unit per round. After the first round, agents re-score the remaining extracted units using the accumulated shared board. Aggregation occurs once, after the final round, and no additional evidence is extracted between rounds.

All schedules achieve 20.0% exact success (one of five instances), CritRec of 1.000, and AF of 80.0%. Mean communication volume is 163 tokens in each condition, but mean total token use rises from 12,168 to 16,778 and 19,163 as re-scoring is added. The two- and three-round schedules therefore use approximately 38% and 57% more total tokens than the one-round schedule. This is a fixed evidence-unit comparison, not a fixed total-token comparison.

The extraction pool also contains at most three units per agent, so each schedule ultimately transmits the same pool, apart from possible ordering differences. Identical observed outcomes are consistent with this constraint; they do not imply that intermediate-round coverage is already complete, or that evidence ordering is universally irrelevant to an LLM aggregator. The sample is too small, and the

Table 44: Fixed-evidence-budget relay pilot on five MuSiQue instances (Qwen3-8B, as identified in the experiment draft). Parentheses give per-round unit allocations. Token counts are means per instance.

Rounds	k	Acc (%)	CritRec	AF (%)	CommTok	TotalTok
1 (3)	3	20.0	1.000	80.0	163	12,168
2 (1 + 2)	3	20.0	1.000	80.0	163	16,778
3 (1 + 1 + 1)	3	20.0	1.000	80.0	163	19,163
Full sharing	—	20.0	—	—	307	1,230

The total budget is three evidence units per agent, not a fixed total-token budget. Later rounds re-score remaining units against the shared board. No new extraction occurs, and the extraction pool also contains at most three units per agent. Full-sharing coverage and AF are omitted because its transmitted-state field was not populated.

setup too constrained, to rule out benefits from new extraction, larger candidate pools, intermediate aggregation, or adaptive evidence queries.

A saved full-sharing pilot row reports 20.0% accuracy, but its audit reads an empty transmitted-state field that the full-sharing implementation does not populate. Its recorded zero CritRec and AF are therefore instrumentation artifacts and are not used as an information-ceiling comparison. The larger attempted multi-round runs record zero total model tokens throughout and are likewise excluded from performance conclusions.

C.4.4 Residual Error Categories and Judge Validation

The classification run uses Qwen3-8B at temperature zero to label saved Qwen-235B predictions, with gold supporting paragraphs supplied to the judge. Candidate IDs are selected from the union of score-ranked and random-relay audit failures, then paired with their full-sharing predictions. This procedure yields 74 candidates, of which 13 have recorded answer F1 equal to 1 under full sharing. We filter out those 13 before computing the category proportions in Table 45. The judge receives the question, supporting paragraphs, prediction, and gold answer.

Table 45: Exploratory error labels for 61 selected non-perfect MuSiQue full-sharing answers (Qwen-235B; Qwen3-8B judge). Thirteen perfect answers are excluded; these labels have no independent human validation.

Category	Count	Share (%)
Reasoning failure	36	59.0
Format / answer mismatch	20	32.8
Partial evidence use	5	8.2
Unclassified	0	0.0

Among the 61 retained cases, the judge assigns 36 (59.0%) to reasoning failure, 20 (32.8%) to format or answer-matching discrepancies, and five (8.2%) to partial evidence use. These exploratory labels motivate scrutiny of both reasoning and answer evaluation. They have no independent human validation, and selection through another protocol's failures prevents interpreting the proportions as unbiased estimates over all full-sharing errors.

Table 46 preserves the original 74-case distribution and shows the effect of the success filter. Ten perfect answers were labelled reasoning failures and three were labelled format failures, illustrating the risk of asking a judge to categorize a prediction under the presupposition that it is wrong. The remaining labels are useful for exploratory diagnosis but still require human checking. The stored answer strings do not include the original reasoning traces, so the judge's format category is interpreted as an answer-string or normalization discrepancy rather than verified evidence of a correct rationale followed by an incorrect output field.

Table 46: Accounting for the aggregation-failure diagnostic sample. The full 74-case label distribution is preserved here; the main analysis uses only the 61 candidates with recorded answer F1 below 1.

Judge category	**All**	**Share (%)**	**F1 = 1**	**F1 < 1**
Reasoning failure	46	62.2	10	36
Format / answer mismatch	23	31.1	3	20
Partial evidence use	5	6.8	0	5
Unclassified	0	0.0	0	0
Total	74	100.0	13	61

Candidate IDs were selected using relay audit failures and then paired with full-sharing predictions. A candidate can therefore have a correct full-sharing answer. Filtering changes the denominator and prevents correct predictions from being counted as failures.

Consistent performance and failure denominators. The original MQ task score is mean answer F1, whereas the supplementary failure analysis uses the binary event that answer F1 is below 1. In the 100-instance Qwen-235B full-sharing run, 35 answers have F1 equal to 1; hence the binary failure rate is 65.0% despite mean F1 of 52.5 points. Table 8 uses this binary rate for comparison with relay AF, rather than the 47.5% complement-of-F1 summary in the original tables. On SB, the task score is binary, so this distinction does not arise.

Table 47: Limitations of the evidence-flow audit framework and associated justifications or mitigations.

Limitation	Details	Justification / Mitigation
Local scoring incompletely captures joint criticality	A fact may score low within one agent's shard but become decisive when combined with evidence from another agent. The observed scores do not establish an impossibility result for all local estimators.	The joint-criticality false-negative analysis (§C.3.1) explicitly quantifies this gap: 77.5% of low-score critical units become false negatives at $k{=}1$. The finding motivates future work on inter-agent criticality estimation.
Gold critical unit approximation from benchmark metadata	We approximate $S^\star$ from gold-annotated support fields (paragraph-level for MuSiQue, fact-level for SiloBench). These may over-count (full paragraph labelled when only one sentence is critical) or miss alternative valid supports.	Paragraph deletion tests the annotated support as a whole, not an isolated atomic fact. Semantic tracking needs validation against independently annotated traces; lexical coverage alone cannot certify sufficiency.
Single-round inference-time coordination only	The audit covers single-round evidence relay and natural-language protocols. Multi-round interaction, dynamic re-querying, and formal privacy or security guarantees are outside the current scope.	The fixed-pool multi-round pilot does not test new evidence acquisition. Targeted re-extraction and aggregation repairs require controlled evaluation before claiming practical repair benefits.

D Limitations and Future Work

The evidence-flow audit framework has three primary limitations, summarised in Table 47. Two immediate validation steps are needed: check critical-fact tracking against independently annotated traces, and compare diagnosis-guided evidence and aggregation repairs against volume-matched additions and generic retries. Gold-assisted restoration would test attribution, while a repair using only original shards would test practical actionability. Neither experiment is reported here. Future work should address the joint-criticality estimation problem directly—perhaps through inter-agent communication of prior-round scores or through multi-shard joint scoring models—and extend the audit to multi-round interaction settings where evidence can be re-queried dynamically. Aggregators trained to reason under partial and noisy evidence (rather than relying on prompt engineering alone) represent a complementary direction; so does extending the framework to agent networks larger than those studied here. On the benchmark side, richer gold annotation at the sentence or clause level (rather than paragraph level) would enable finer-grained audit metrics and reduce the approximation error in critical unit identification.