CONSENSUS IS NOT EVIDENCE: EVIDENCE LINEAGE FOR REPLICATION-SAFE MULTI-AGENT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Many multi-agent LLM systems treat repeated agreement as additional support. When reports share upstream evidence, this can count one observation multiple times and suppress a minority report grounded in new evidence. We formalize *evidence lineage* by grouping traces by answer-relevant upstream observations and show that marginal accuracy and pairwise correlation cannot identify consensus reliability. Even identical answers and calibrated reliabilities can require opposite Bayes decisions depending on whether reports are independent or replicated. This motivates *replication invariance*: adding a non-informative replica should change neither prediction nor confidence. RIDE operationalizes these principles through an acquire–group–aggregate protocol that audits lineage structure, acquires distinct upstream evidence when needed, collapses exact replicas, and aggregates within and across lineages without adding mass to duplicates. The framework separates failures to obtain new evidence from failures to preserve its influence during aggregation. Controlled interventions isolate answer-bearing support as the driver of acquisition and show that handling must preserve an informative lineage against redundant votes. In a non-oracle study with parameters fixed before evaluation, RIDE-CAL improves over the best fixed source by 7.5 points and reduces binary answer NLL from 1.646 to 0.650. Across 3,600 exact-replication attacks, its replication-invariant design yields a 0% prediction-or-confidence change rate versus 69.08% for confidence weighting. Replication safety persists under provenance corruption, while false merges and aggressive overlap grouping emerge as accuracy risks. Reliable multi-agent scaling should therefore optimize and audit evidence lineages, not raw response count.

1 INTRODUCTION

Voting, self-consistency, and multi-agent debate are widely used to improve language-model reliability (Wang et al., 2023; Du et al., 2023; Chan et al., 2023; Chen et al., 2024; Li et al., 2024). These methods often treat each response as an additional observation, so more agreeing instances receive more decision mass.

Consider three agents that repeat one answer from the same incomplete context and a fourth that reaches a different answer after observing a missing supporting passage. Instance voting favors the majority even though the minority report contains the only new answer-relevant observation.

This example distinguishes report multiplicity from evidence multiplicity. Because agents may share retrieval results, tools, or upstream messages, five agreeing reports may reflect five distinct observations or one repeated observation. We call the answer-relevant provenance underlying a trace its evidence lineage (Figure 1).

Response counting resembles the Condorcet jury theorem: independently correct judgments above chance make majority accuracy approach one as the group grows (de Condorcet, 1785). In LLM systems, however, the unit of independent evidence is rarely observed.

Report provenance reveals which outputs reuse the same observation and which contribute distinct evidence. It also separates failures to acquire evidence from failures to preserve its influence during

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.

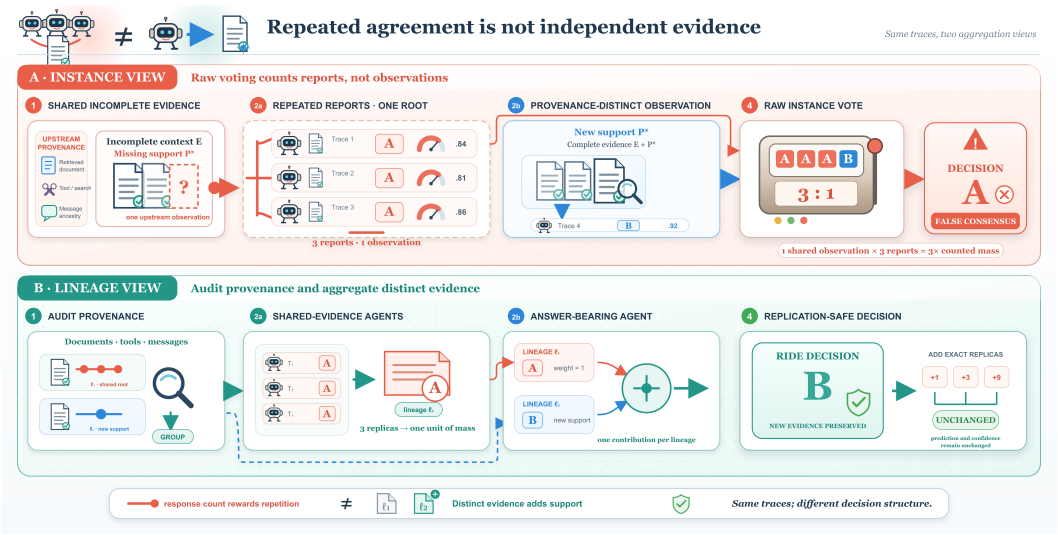

Figure 1: Instance voting gives three replicas of incomplete evidence more mass than one provenance-distinct report (top). RIDE audits ancestry, collapses exact replicas, and aggregates by lineage so replication adds no mass while new evidence remains available (bottom). Values are illustrative.

aggregation. Correlation alone cannot recover this structure. This is a problem-formulation error, not merely a weakness of majority voting: a generally valid aggregation rule must know which reports constitute new observations.

We formalize this distinction through evidence lineage. Under exchangeable dependence, identical marginal accuracy and pairwise correlation can conceal different majority-error floors. More strongly, without dependence information, identical answers and confidences can require opposite Bayes decisions. We then characterize exactly when aggregation is immune to replication and decompose system regret into information left unacquired and distortion introduced during handling.

Replication-Invariant Deliberation through Epistemic Interventions (RIDE) turns these results into an acquire–group–aggregate protocol. We evaluate two instantiations. RIDE-CONTROLLED uses known provenance and a designated source to isolate acquisition and handling. RIDE-CAL learns candidate reliability, aggregates answers within and across lineages, discounts overlap, calibrates confidence, and supports sequential acquisition. RIDE-CAL is frozen: its parameters and thresholds are selected on development data and fixed before the 200-task held-out evaluation.

Three complementary studies evaluate the framework. A 2×2 design isolates answer-bearing support, a held-out non-oracle study compares aggregation alternatives, and provenance corruptions test false merges, false splits, missing identifiers, partial overlap, and adversarial splits.

Our contributions are:

- We formulate answer-relevant *evidence lineage* as an explicit decision variable and prove that accuracy, pairwise correlation, answers, and calibrated reliabilities are insufficient without report-dependence information;
- We characterize replication invariance exactly through replica-equivalence classes and decompose excess log loss into acquisition and handling regret, separating missing information from distortion after acquisition;
- We identify answer-bearing support as the acquisition driver, demonstrate replication-safe non-oracle aggregation with substantially improved calibration, and map its failure surface under imperfect provenance.

2 Related Work

Multi-agent scaling and correlated judgments. The benefits of multi-agent aggregation depend on how much distinct information agents contribute. Self-consistency and debate aggregate sampled paths or interacting outputs (Wang et al., 2023; Du et al., 2023; Chan et al., 2023; Chen et al., 2024), yet errors remain correlated across providers (Kim et al., 2025). Nine LLM judges may provide only about two effective votes (Kohli, 2026), while limited attention and model similarity can reduce collective information and bias oversight (Liu et al., 2026; Goel et al., 2025). Representing lineage explicitly exposes the gap between panel size and informational diversity.

Source copying and dependence-aware aggregation. Dependence-aware aggregation spans correlated-vote theory (Ladha, 1992), removal of repeated information from known communication graphs (Hamdi & Krishnamurthy, 2013), and estimation of source copying and accuracy (Dong et al., 2009). Other work addresses unknown correlation (de Oliveira et al., 2021) or estimates complementarity and candidate quality (Turkmen et al., 2026; Jiang et al., 2023a; Cobbe et al., 2021). RIDE-CAL combines learned reliability with a structural constraint that gives replicas no additional decision mass.

Communication, calibration, and closed systems. Communication can homogenize outputs (Liu, 2026), and message genealogy can expose error cascades (Xie et al., 2026). CAGE-CAL calibrates communication-induced agreement using counterfactual graphs (Huang et al., 2026). Decision-theoretic and data-processing analyses further bound networks that only transform fixed evidence (Ao et al., 2026; Shin, 2026). These works study propagation, compression, or calibration within an information state; our controlled intervention separates such handling from acquiring a conditionally informative lineage.

Evidence acquisition and active retrieval. Retrieval-augmented generation supplies documents (Lewis et al., 2020), active retrieval decides when to obtain more (Jiang et al., 2023b), and research-and-revision retrieves corrective support (Gao et al., 2023a). Information-bottleneck and informational-substitutes results characterize bandwidth and diminishing returns (Yu et al., 2026; Chen & Waggoner, 2016; Nemhauser et al., 1978). Our decomposition distinguishes information left unacquired from distortion after acquisition.

Provenance, lineage, and dependent reports. Database provenance records why and where outputs arise (Buneman et al., 2001; Cheney et al., 2009), citation-grounded generation connects claims to documents (Gao et al., 2023b), and peer prediction elicits correlated private reports from strategic agents (Miller et al., 2005; Prelec, 2004). Evidence lineage groups document, tool, and message ancestry at an answer-relevant resolution for non-strategic LLM traces, making merge, split, missing-identifier, and overlap errors explicit decision inputs.

3 Why Consensus Requires Provenance

Correlation can reveal dependence without identifying its consequences. We show why evidence provenance is required to distinguish independent from replicated support in multi-agent systems, characterize replication-safe aggregation, and separate acquisition from handling risk.

3.1 Evidence Lineages and the Limits of Correlation

Let $Y \in \{-1, +1\}$ be the correct answer and $X_i \in \{0, 1\}$ indicate whether agent i is correct. A trace is

$$T_i = (A_i, P_i, E_i, L_i), \tag{1}$$

where A_i is the answer, P_i is a confidence distribution, E_i is cited evidence or claims, and L_i is a lineage label derived from evidence, tool, and message ancestry. Raw provenance additionally includes model and prompt metadata; L_i coarsens this record at the chosen answer-relevant resolution, so model or prompt changes alone need not create a new lineage.

Definition 1 (Evidence lineage)**.** *At a chosen provenance resolution, an evidence lineage ℓ is the equivalence class of traces generated from the same answer-relevant upstream evidence state E_ℓ.*

Different models or messages remain in one lineage whenever they transform the same evidence state; a trace enters a new lineage only when its evidence state contains an answer-relevant observation that the other traces did not access.

Lineage is therefore relative to the information boundary under study. In our controlled experiment, every trace that observes the same context with a critical paragraph withheld belongs to the shared lineage, while the trace that observes the complete support belongs to a second lineage.

A concrete HotpotQA instance follows this pattern: three traces from an incomplete shared context favor 2004, while a trace that observes the withheld sentence correctly returns 1974 (Appendix D.3).

We use a de Finetti-style exchangeable model (de Finetti, 1937; Ladha, 1992). Let $\Theta \in [0, 1]$ be the latent competence state of the team on a task instance. Conditional on Θ, correctness variables are independent:

$$X_i \mid \Theta \overset{\text{iid}}{\sim} \text{Bernoulli}(\Theta), \qquad \mathbb{E}[\Theta] > \tfrac{1}{2}. \tag{2}$$

The marginal assumption says that an agent is better than chance on average. It does not rule out shared task states in which all agents are more likely wrong than correct.

Theorem 1 (Correlated jury error floor)**.** *For odd n, let $M_n = \mathbf{1}[\sum_{i=1}^{n} X_i < n/2]$. If $\mathbb{P}(\Theta = 1/2) = 0$, then*

$$\lim_{n\to\infty} \mathbb{P}(M_n = 1) = \mathbb{P}(\Theta < 1/2). \tag{3}$$

Therefore, if $\mathbb{P}(\Theta < 1/2) > 0$, majority error does not vanish as the number of agents grows.

Proof sketch. The strong law gives $n^{-1} \sum_i X_i \to \Theta$ almost surely, after which bounded convergence yields the result. When $\Theta = p > 1/2$ is constant, Hoeffding's inequality instead gives exponentially decreasing error (Hoeffding, 1963). Full proofs appear in Appendix B.

The floor depends on the probability mass below $1/2$, not on correlation alone.

Proposition 1 (Correlation does not identify consensus reliability)**.** *Marginal accuracy and pairwise correctness correlation do not identify the limit in Theorem 1. Two explicit mixing laws have the same accuracy 0.6 and correlation $1/6$ but majority-error floors $1/2$ and $1/6$ (Appendix B).*

Thus correlation diagnoses dependence but cannot determine whether agreement represents independent confirmation or repetition of a shared failure state.

3.2 Provenance Necessity and Replication-Invariant Aggregation

The problem runs deeper than correlation insufficiency: exact answers and calibrated reliabilities are still insufficient without report-dependence information.

Theorem 2 (Provenance non-identifiability)**.** *No aggregator that observes only report answers and calibrated reliabilities can be Bayes-optimal across all dependence structures compatible with those observations. For reports $\{(+1, .75)^3, (-1, .9)\}$, the Bayes decision is $+1$ when the positive reports are independent and -1 when they replicate one signal.*

Intuitively, three independent 75% reports outweigh a 90% dissenter; three replicas do not. Evidence lineage supplies the missing dependence information; model identity or surface diversity alone need not. Appendix B gives the log-odds calculation.

Let $\mathcal{A}(S)$ return a decision and confidence from a multiset of traces S. A controlled duplication operation creates T' from T without a new model call, observation, computation, or message and preserves the answer, confidence, claims, and answer-relevant evidence. Let $T \approx T'$ denote membership in the equivalence closure of this duplication relation. Traces may share an evidence lineage without being replica-equivalent: independently sampled or differently computed traces from one evidence state can disagree and remain distinct inputs to aggregation.

Definition 2 (Replication invariance)**.** *An aggregator is replication invariant if, for every transcript S and $T' \approx T$,*

$$\mathcal{A}(S \cup \{T\}) = \mathcal{A}(S \cup \{T, T'\}). \tag{4}$$

Equality applies to both prediction and confidence.

Let $q_{\approx}(S)$ be the set of replica-equivalence classes represented in S.

Theorem 3 (Replica-quotient characterization)**.** *An aggregator is replication invariant if and only if it factors as $\mathcal{A}(S) = \bar{\mathcal{A}}(q_{\approx}(S))$ for a function over represented replica-equivalence classes.*

This characterization establishes replica quotienting as the structural design principle for replication-safe aggregation.

Replication invariance therefore removes multiplicity only within replica classes. An evidence lineage may contain several non-replica classes whose processing variation remains available to a within-lineage summarizer. A hierarchical aggregator can take the form $\mathcal{A}(S) = h(\{g_\ell(q_{\approx}(S_\ell))\}_\ell)$, where g_ℓ summarizes one lineage and h combines lineages. Additive raw-instance rules remain manipulable by sufficiently many replicas of an opposing trace (Appendix B).

3.3 Acquisition and Handling Regret

Let V index all available lineages, $Q \subseteq V$ those acquired, $p_Q = p(Y \mid S, T_Q)$, $\mathcal{R}(q, Q)$ expected log loss, and $\mathcal{R}^\star(V) = H(Y \mid S, T_V)$ the full-information Bayes risk.

Theorem 4 (Acquisition–handling decomposition)**.** *The excess log loss relative to a Bayes predictor with all available lineages decomposes as*

$$\mathcal{R}(q_{\mathcal{A}}, Q) - \mathcal{R}^\star(V) = \underbrace{I(Y; T_{V\setminus Q} \mid S, T_Q)}_{\textit{acquisition regret}} + \underbrace{\mathbb{E}[D_{\mathrm{KL}}(p_Q \| q_{\mathcal{A}})]}_{\textit{handling regret}}. \tag{5}$$

The first term is information left outside the transcript; the second is distortion introduced after acquisition. This specializes posterior-distortion analyses of common-evidence delegation (Ao et al., 2026) to omitted lineages and non-Bayes handling. Acquiring all lineages removes the first, while Bayes-optimal handling removes the second.

Remark (closed-transcript limitation). Under the Markov condition $Y \to (S, T_Q) \to T$, discussion generated solely from the acquired transcript satisfies $I(Y; T \mid S, T_Q) = 0$ and cannot introduce a new exogenous observation (Ao et al., 2026; Shin, 2026), although it may improve computation over information already present. We test this boundary by withholding a uniquely answer-bearing paragraph from every debating agent.

3.4 RIDE: Framework and Instantiations

RIDE is the general acquire–group–aggregate framework: it identifies lineage structure, diagnoses insufficient support, acquires sources, and aggregates after collapsing exact replicas. RIDE-CONTROLLED uses known provenance to isolate acquisition and handling; RIDE-CAL is the frozen non-oracle instantiation. These components target the two terms of Theorem 4, respectively.

Controlled mechanism rule. Let R denote retained support, D distractors, and $P^\star$ uniquely answer-bearing support. RIDE-CONTROLLED gives shared agents $R \cup D$ and a designated independent source $R \cup P^\star$. When they conflict, it routes to the designated source. Comparing the same solver with and without $P^\star$ identifies acquisition; comparing RIDE-CONTROLLED with instance voting on the same traces identifies whether redundant instances erase an informative source.

Non-oracle aggregation. RIDE-CAL receives retrieval lineages without knowing which is informative. It first collapses exact replicas. For each answer a in lineage ℓ, a development-fitted logistic model estimates reliability $r_{\ell a}$ from source identity, confidence, within-lineage agreement, support score, retrieval scores, context length, cross-lineage answer support, and document overlap. A Platt map then calibrates the estimate. RIDE-CAL preserves within-lineage disagreement through

$$\pi_\ell(a) = \mathrm{softmax}_a(\mathrm{logit}(r_{\ell a})/\tau)\,, \tag{6}$$

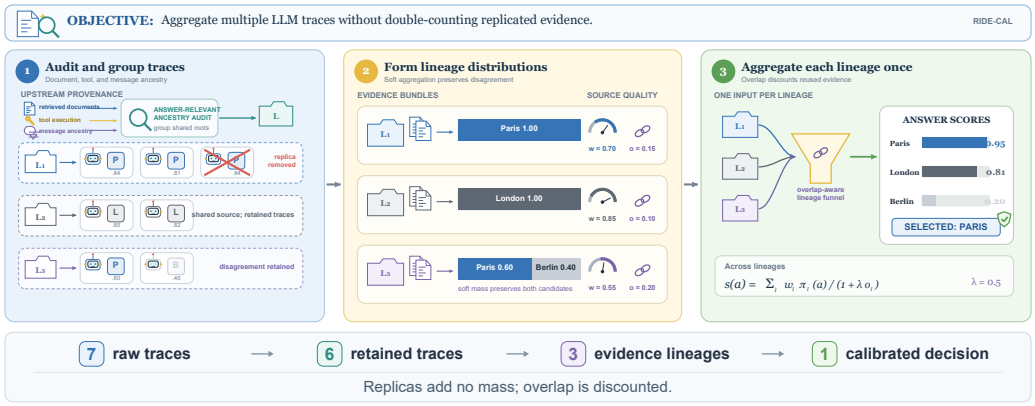

Figure 2: **Hierarchical replication-invariant aggregation.** Exact replicas add no mass; π_ℓ preserves non-replica disagreement before reliability- and overlap-weighted lineage aggregation. Values are illustrative.

rather than selecting one answer per lineage. With $w_\ell = \max_a r_{\ell a}$ and overlap summary o_ℓ, the cross-lineage score is

$$s(a) = \sum_\ell \frac{w_\ell \pi_\ell(a)}{1 + \lambda o_\ell}. \tag{7}$$

The winning score share is task-calibrated by a second Platt map.

The candidate model, τ, λ, and calibration maps are selected by nested cross-validation and frozen before confirmation. An active variant acquires lexical, complementary-PRF, and dense sources sequentially, stopping on reliability, entropy, answer-margin, and calibrated-confidence thresholds. This instantiation turns source diversity, within-lineage disagreement, and document overlap into a deployable decision rule. Appendix C.1 gives pseudocode for frozen source construction and hierarchical inference; Figure 2 summarizes the hierarchy.

4 Experimental Design

Our evaluation proceeds in four stages: a controlled mechanism study using RIDE-CONTROLLED, a support–distractor factorial, a frozen non-oracle confirmation, and a provenance-sensitivity analysis.

Controlled mechanism study. The controlled study uses 159 MuSiQue questions (Trivedi et al., 2022); a frozen extension uses 100 HotpotQA (Yang et al., 2018) and 100 GSM8K questions (Cobbe et al., 2021). Shared document agents lack the uniquely answer-bearing support, whereas the designated document source receives complete gold support; the GSM8K source is a deterministically verified calculator result. These RIDE-CONTROLLED contrasts identify evidence acquisition and same-trace handling but use oracle source roles. Appendix D reports the complete condition-level results.

Support-by-distractor factorial. To remove the document-regime confound, we cross answer-bearing support $P^\star \in \{0, 1\}$ with original distractors $D \in \{0, 1\}$. Every task therefore has R, $R \cup D$, $R \cup P^\star$, and $R \cup D \cup P^\star$ under the same Qwen model, temperature, output cap, and prompt. The paired sample contains all 159 MuSiQue and 100 HotpotQA tasks, 777 newly collected traces, and no exclusions. This design separately estimates the support effect at each distractor level and the distractor effect at each support level.

Frozen non-oracle confirmation. We freeze 200 held-out HotpotQA tasks and retrieve lexical BM25, dense, and complementary-PRF lineages. Each supplies four documents and two Qwen traces; no method knows which source is correct. Complementary PRF is designed to prevent

construction-time identity without assuming semantic independence. It replaces one document only when its top-four set duplicates the lexical top-four set. Nested cross-validation on 99 disjoint development tasks selects RIDE-CAL and active-acquisition parameters before the 1,200 confirmation calls. Baselines cover fixed sources, instance and lineage aggregation, hard and verifier selectors, exact-replica collapse, and an oracle ceiling.

Imperfect provenance. Using the same 200 tasks and 1,200 traces, we perturb labels without new calls. Across 100 seeds per level, false merges combine sources, false splits divide one source, and adversarial splits isolate disagreements. Missing-provenance conditions remove document identifiers; an overlap graph groups lineages across Jaccard thresholds. Trace contents and the frozen selector never change.

Metrics and audit. Accuracy uses normalized exact match, with token F1 secondary. Binary answer NLL evaluates the reported probability that the selected answer is correct, and ECE uses ten bins. A source's representative answer is selected by within-source agreement, with confidence breaking ties. Conflict tasks have at least two disagreeing source representatives; wrong consensus means that all panel answers are identical and incorrect, for which we report both count and denominator.

We use 10,000 paired bootstrap samples and exact McNemar tests. The replication audit adds 1, 3, or 9 copies of each of six traces across 200 tasks, producing $200 \times 6 \times 3 = 3{,}600$ attacks per method. Integrity checks verify task freezes, trace completeness, evidence assignments, and selector immutability.

5 RESULTS

5.1 ANSWER-BEARING SUPPORT DRIVES THE ACQUISITION EFFECT

Table 1 reports the four support–distractor cells and the paired contrasts separating acquisition from distraction.

Table 1: Factorial accuracy (%) and paired effects. $P^\star$ is the uniquely answer-bearing support and D the original distractors; for example, $\Delta P^\star \mid \neg D$ is the support effect when distractors are absent.

Dataset	R	$R{+}D$	$R{+}P^\star$	$R{+}D{+}P^\star$	$\Delta P^\star \mid \neg D$	$\Delta P^\star \mid D$	$\Delta D \mid P^\star$
MuSiQue	7.55	4.40	47.80	47.17	+40.25	+42.77	-0.63
HotpotQA	16.00	15.00	59.00	49.00	+43.00	+34.00	-10.00

Across both datasets, answer-bearing support produces the dominant accuracy gain. On MuSiQue, adding $P^\star$ improves accuracy by 40.25 points without distractors (95% CI: 32.70–48.43) and 42.77 points with them (34.59–50.94); distractor effects are small. On HotpotQA, support adds 43 points without distractors (33–53) and 34 with them (24–44). Distractors have a smaller negative effect when support is present (-10 points, 95% CI: -18 to -3). Thus, answer-bearing evidence drives acquisition gains, with a secondary dataset-specific distractor effect.

Having isolated acquisition under controlled interventions, we evaluate lineage-aware aggregation without a designated correct source.

5.2 FROZEN NON-ORACLE AGGREGATION

Table 2 compares the frozen non-oracle methods on accuracy, calibration, and source use.

Table 2: Frozen 200-task confirmation. Accuracies are percentages; the oracle is a ceiling, not a deployable baseline.

Method	Acc.	Conflict acc.	NLL	ECE	Sources
Best fixed source	44.5	29.6	3.670	0.390	1.00
Confidence-weighted instances	50.5	39.1	1.646	0.216	3.00
Reliability-weighted lineage vote	51.0	40.0	3.330	0.387	3.00
Hard lineage selector	51.0	40.0	0.703	0.120	3.00
Verifier selector	40.0	25.2	1.615	0.341	3.00
RIDE-CAL	**52.0**	**40.9**	**0.650**	0.089	3.00
Active RIDE-CAL	**53.0**	**42.6**	0.662	**0.081**	1.86
Oracle best available	63.5	59.1	0.292	0.188	3.00

Among non-active methods, RIDE-CAL attains the highest overall and conflict-subset accuracies: 52.0% and 40.9%. It improves over the best fixed source by 7.5 points (95% CI: 3–12; $p = 0.0026$) and exceeds confidence-weighted instances by 1.5 points overall and 1.74 points on the 115-task conflict subset. The latter intervals cross zero (-0.5 to 4.0 overall; -1.74 to 5.22 on conflicts), making the accuracy advantage directional at this sample size.

The larger separation appears in calibration. RIDE-CAL lowers binary answer NLL from 1.646 to 0.650, a paired difference of -0.996 (95% CI: -1.493 to -0.546), and ECE from 0.216 to 0.089. Active RIDE-CAL reaches 53.0% while acquiring only 1.86 of three sources on average. Thus lineage-aware source accounting improves confidence quality while supporting adaptive evidence budgets.

The calibration gain is concentrated on the 85 non-conflict tasks, where NLL falls from 2.991 to 0.637. On conflict tasks, RIDE-CAL and confidence weighting have similar NLLs of 0.659 and 0.651. The method therefore corrects a particularly consequential failure mode: overconfidence in apparently unanimous panels, including unanimously wrong ones.

Exact replication yields the clearest structural result: across 3,600 attacks, confidence-weighted instances change prediction or confidence in 69.08% of cases, while RIDE-CAL changes in 0%.

5.3 IMPERFECT PROVENANCE SEPARATES SAFETY FROM ACCURACY

Figure 3 varies only lineage metadata, separating semantic accuracy risks from replication-safety failures.

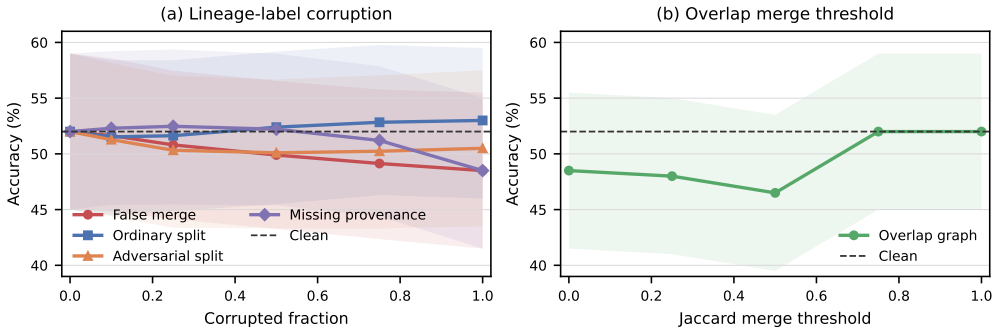

Figure 3: Frozen provenance perturbations for RIDE-CAL. Label corruptions vary the affected fraction; overlap grouping varies the Jaccard threshold. Bands are 95% intervals over tasks and 100 perturbation seeds. Exact-replication change remains 0% throughout.

Replication safety holds under clean provenance; the remaining deployment question is how errors in lineage construction affect accuracy and calibration. Clean RIDE-CAL accuracy is 52.0%. Merging all source lineages, or removing all document provenance (which collapses every trace into a single lineage), lowers accuracy to 48.5%. Ordinary false splitting does not reduce accuracy (53.0% at full corruption) and leaves NLL effectively unchanged (0.650 to 0.657); adversarially isolating

disagreeing answers lowers accuracy to 50.5%. Overlap clustering is most brittle when aggressive: threshold 0.5 reaches 46.5%, whereas a conservative 0.75 threshold recovers 52.0%. Replication change remains 0% in every condition, even when grouping errors reduce accuracy.

5.4 Controlled Diagnostics Isolate Handling Effects

Having isolated acquisition, we compare aggregation rules on identical controlled trace multisets to test the complementary mechanism: handling. The designated-source rule beats instance voting by 48.43 points on MuSiQue, 28 points on HotpotQA, and 33 points when preserving a verified GSM8K calculator result. Fixing the informative source makes these clean mechanism contrasts. Crossing the evidence boundary by acquiring $P^\star$ drives the improvement. Closed-room debate and higher sampling temperature reorganize traces within the fixed withheld-support evidence state.

6 Discussion

Structural safety and semantic correctness. Replication invariance prevents exact copies from gaining decision mass, while source reliability and lineage inference determine which distinct observations to trust. The frozen confirmation shows 0% replication changes for RIDE-CAL versus 69.08% for confidence weighting; the 63.5% oracle ceiling marks remaining source-selection headroom.

From controlled labels to open systems. Practical systems can infer lineage from retrieved-document identifiers, citations, tool paths, and message ancestry. Our perturbations show that false merges and aggressive overlap clustering suppress distinct evidence, while ordinary false splits are comparatively benign; conservative grouping is therefore preferable when provenance is uncertain.

Implications for benchmarks and evaluation. Multi-agent benchmarks should record answer-relevant provenance and report effective lineage counts, cross-lineage conflicts, acquisition cost, and stability under replication. This prevents agent count or surface diversity from masquerading as independent confirmation and keeps acquisition, handling, calibration, and provenance error separately auditable.

7 Limitations

Our evaluation covers multi-hop document QA and verified computational reasoning; broader model families and less structured environments may require additional study. Lineage currently uses auditable document identifiers, tool paths, and message ancestry, while active acquisition trades retrieval and inference cost for adaptive evidence use. Deployment-specific work can extend lineage estimation and optimize this tradeoff under explicit budgets.

8 Conclusion

Consensus is informative only when it reflects new evidence. Correlation can diagnose dependence without identifying its decision consequences, and lineage-blind aggregation cannot generally distinguish independent support from replication. Evidence lineage makes that missing structure explicit. Theorem 3 characterizes exactly when replication cannot change a prediction or confidence, while Theorem 4 separates information absent from the transcript from distortion introduced after acquisition.

Three empirical results reinforce this framework. Factorial interventions identify answer-bearing support as the driver of acquisition, with smaller distractor effects. Frozen non-oracle RIDE-CAL improves accuracy and calibration while remaining unchanged under exact replication. Provenance corruption identifies false merging and aggressive overlap grouping as the primary risks to source-aware aggregation.

Together, the results shift the unit of multi-agent scaling from responses to evidence. Future systems should acquire conditionally new observations, record their provenance, and aggregate over informative lineages rather than increase agent count alone.

AI USE STATEMENT

Generative AI tools were used solely to assist with translation, software code generation, and language editing. All AI-generated code and text were reviewed, revised, and validated by the authors. The research questions, theoretical development, experimental design, result interpretation, and scientific conclusions were developed independently by the authors. The authors take responsibility for the final content of this work.

ETHICS STATEMENT

Multi-agent agreement can create unwarranted confidence in high-stakes applications. Our findings argue for provenance logging and explicit evidence lineage rather than presenting repeated model outputs as independent validation. The experiments use public benchmark data and model APIs and contain no human-subject data. We retain raw responses, model identifiers, and analysis artifacts to support reproducibility.

REPRODUCIBILITY STATEMENT

Appendix B states the assumptions and provides complete proofs for the theoretical results. Appendix C documents frozen source construction and hierarchical inference in pseudocode, reports selector and acquisition settings, and provides full factorial, non-oracle, and provenance-perturbation results. Appendix D gives complete controlled-study conditions and an illustrative evidence intervention. The experiment artifacts retain task manifests, frozen configurations, raw model responses, evaluation outputs, and versioned records used to generate the reported statistics.

REFERENCES

Ruicheng Ao, Siyang Gao, and David Simchi-Levi. On the reliability limits of LLM-based multi-agent planning. *arXiv preprint arXiv:2603.26993*, 2026. URL https://arxiv.org/abs/2603.26993.

Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and where: A characterization of data provenance. In *Database Theory—ICDT 2001*, pp. 316–330, 2001. doi: 10.1007/3-540-44503-X_20. URL https://doi.org/10.1007/3-540-44503-X_20.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023. URL https://arxiv.org/abs/2308.07201.

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

Yiling Chen and Bo Waggoner. Informational substitutes. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science*, pp. 239–247, 2016. doi: 10.1109/FOCS.2016.33. URL https://arxiv.org/abs/1703.08636.

James Cheney, Laura Chiticariu, and Wang-Chiew Tan. Provenance in databases: Why, how, and where. *Foundations and Trends in Databases*, 1(4):379–474, 2009. doi: 10.1561/1900000006. URL https://doi.org/10.1561/1900000006.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Rei Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Nicolas de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris, 1785.

Bruno de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7(1):1–68, 1937.

Henrique de Oliveira, Yuhta Ishii, and Xiao Lin. Robust aggregation of correlated information. *arXiv preprint arXiv:2106.00088*, 2021. URL https://arxiv.org/abs/2106.00088.

Xin Luna Dong, Laure Berti-Équille, and Divesh Srivastava. Integrating conflicting data: The role of source dependence. *Proceedings of the VLDB Endowment*, 2(1):550–561, 2009. doi: 10.14778/1687627.1687690. URL https://doi.org/10.14778/1687627.1687690.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023. URL https://arxiv.org/abs/2305.14325.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, 2023a. doi: 10.18653/v1/2023.acl-long.910. URL https://aclanthology.org/2023.acl-long.910/.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023b. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Shashwat Goel, Joschka Struber, Ilze Amanda Auzina, Karuna K. Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines AI oversight. *arXiv preprint arXiv:2502.04313*, 2025. URL https://arxiv.org/abs/2502.04313.

Maziyar Hamdi and Vikram Krishnamurthy. Removal of data incest in multi-agent social learning in social networks. *arXiv preprint arXiv:1309.6687*, 2013. URL https://arxiv.org/abs/1309.6687.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Jiatan Huang, Mingchen Li, Ziming Li, Sunjae Kwon, Hong Yu, and Chuxu Zhang. Counterfactual graph for multi-agent LLM calibration. *arXiv preprint arXiv:2605.30653*, 2026. URL https://arxiv.org/abs/2605.30653.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, 2023a. doi: 10.18653/v1/2023.acl-long.792. URL https://aclanthology.org/2023.acl-long.792/.

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

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. *arXiv preprint arXiv:2506.07962*, 2025. URL https://arxiv.org/abs/2506.07962.

Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels. *arXiv preprint arXiv:2605.29800*, 2026. URL https://arxiv.org/abs/2605.29800.

Krishna K. Ladha. The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 36(3):617–634, 1992. doi: 10.2307/2111584.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *arXiv preprint arXiv:2402.05120*, 2024. URL https://arxiv.org/abs/2402.05120.

Kaixuan Liu, Guojun Xiong, Weinan Zhang, and Shengpu Tang. Social networks of LLM agents. *arXiv preprint arXiv:2607.03695*, 2026. URL https://arxiv.org/abs/2607.03695.

Zewen Liu. BOUNDARY_SYNC: Measuring communication-induced representational coupling in multi-agent LLM systems. *arXiv preprint arXiv:2607.01600*, 2026. URL https://arxiv.org/abs/2607.01600.

Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005. doi: 10.1287/mnsc.1050.0379. URL https://doi.org/10.1287/mnsc.1050.0379.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14:265–294, 1978. doi: 10.1007/BF01588971.

Dražen Prelec. A bayesian truth serum for subjective data. *Science*, 306(5695):462–466, 2004. doi: 10.1126/science.1102081. URL https://doi.org/10.1126/science.1102081.

Kwan Soo Shin. The reasoning trap: An information-theoretic bound on closed-system multi-step LLM reasoning. *arXiv preprint arXiv:2605.01704*, 2026. URL https://arxiv.org/abs/2605.01704.

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

Yigit Turkmen, Baturalp Buyukates, and Melih Bastopcu. Don't always pick the highest-performing model: An information theoretic view of LLM ensemble selection. *arXiv preprint arXiv:2602.08003*, 2026. URL https://arxiv.org/abs/2602.08003.

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

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

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

Wendi Yu, Lianhao Zhou, Xiangjue Dong, Sai Sudarshan Barath, Declan Staunton, Byung-Jun Yoon, Xiaoning Qian, James Caverlee, and Shuiwang Ji. When do multi-agent systems help? an information bottleneck perspective. *arXiv preprint arXiv:2607.16133*, 2026. URL https://arxiv.org/abs/2607.16133.

A NOTATION AND TERMINOLOGY

Table 3: Core notation, terminology, and system variants.

Symbol or term	Meaning
Evidence lineage ℓ	Traces generated from the same answer-relevant upstream evidence state at the chosen provenance resolution.
Replica equivalence $T \approx T'$	A controlled duplication relation in which T' adds no new observation, computation, or message beyond T.
$q_{\approx}(S)$	The replica-equivalence classes represented in transcript S.
Replication invariance	Adding a replica changes neither the prediction nor its confidence.
R, D, $P^\star$	Retained support, distractors, and uniquely answer-bearing support in the controlled factorial design.
Acquisition regret	Information about Y that remains in available but unacquired lineages.
Handling regret	Posterior distortion introduced after the acquired traces enter the transcript.
Frozen	Source construction, learned parameters, calibration maps, and thresholds are selected on development data and fixed before confirmation.
Non-oracle	Inference proceeds without a designated correct source or privileged source label.
RIDE, RIDE-CONTROLLED, RIDE-CAL	The general framework, its controlled mechanism instantiation, and its frozen non-oracle instantiation, respectively.

B PROOFS

B.1 CORRELATED JURY ERROR FLOOR AND CORRELATION INSUFFICIENCY

Let Θ be defined on a probability space and let $X_1, X_2, \ldots$ be conditionally i.i.d. Bernoulli(Θ). Conditional on $\Theta = \theta$, the strong law of large numbers gives

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \theta \quad \text{almost surely.} \tag{8}$$

If $\theta < 1/2$, there exists a finite random N such that the sample mean is below $1/2$ for all $n > N$, so the majority is wrong. If $\theta > 1/2$, the majority is eventually correct. Since $\mathbb{P}(\Theta = 1/2) = 0$,

$$M_n \to \mathbf{1}[\Theta < 1/2] \quad \text{almost surely.} \tag{9}$$

Because $0 \le M_n \le 1$, bounded convergence yields

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}(M_n = 1) &= \mathbb{E}[\mathbf{1}[\Theta < 1/2]] \\ &= \mathbb{P}(\Theta < 1/2). \end{aligned} \tag{10}$$

When $\Theta = p > 1/2$ is constant, conditional and marginal independence coincide. Hoeffding's inequality gives

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{n}\sum_i X_i \le \frac{1}{2}\right) \\ &\qquad \le \exp\left[-2n\left(p - \frac{1}{2}\right)^2\right]. \end{aligned} \tag{11}$$

Thus the non-vanishing floor comes from variation in the shared latent state, not from weak marginal agents.

For Proposition 1, let $p = \mathbb{E}[\Theta]$. Marginally, X_i is Bernoulli(p), while conditional independence gives

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= \mathbb{E}[\Theta^2] - \mathbb{E}[\Theta]^2 = \mathrm{Var}(\Theta), \\ \rho &= \frac{\mathrm{Var}(\Theta)}{p(1-p)}. \end{aligned} \tag{12}$$

Let $\mathbb{P}(\Theta_A = 0.4) = \mathbb{P}(\Theta_A = 0.8) = 1/2$. Then $\mathbb{E}[\Theta_A] = 0.6$ and $\mathbb{E}[\Theta_A^2] = 0.4$, so its variance is 0.04. For Θ_B,

$$\begin{aligned} \mathbb{E}[\Theta_B] &= \tfrac{1}{6}(0.2) + \tfrac{1}{2}(0.6) + \tfrac{1}{3}(0.8) = 0.6, \\ \mathbb{E}[\Theta_B^2] &= \tfrac{1}{6}(0.04) + \tfrac{1}{2}(0.36) + \tfrac{1}{3}(0.64) = 0.4. \end{aligned} \tag{13}$$

It also has variance 0.04, hence both models have $\rho = 0.04/(0.6\cdot 0.4) = 1/6$. Theorem 1 nevertheless gives limits $1/2$ and $1/6$, the respective masses below $1/2$.

B.2 Provenance Non-identifiability

Assume a uniform prior on $Y \in \{-1, +1\}$ and symmetric binary signals. A positive report with reliability r contributes likelihood ratio $r/(1-r)$ in favor of $Y = +1$; a negative report contributes $(1-r)/r$. If the three 0.75 positive reports in $\mathcal{O}$ are conditionally independent, the posterior odds are

$$\left(\frac{0.75}{0.25}\right)^3 \left(\frac{0.1}{0.9}\right) = 3^3/9 = 3, \tag{14}$$

so the posterior log odds are $\log 3$ and the Bayes decision is $+1$.

In the replicated world, the three positive reports are deterministic copies of one 0.75 signal. They contribute one likelihood factor rather than three, giving

$$\left(\frac{0.75}{0.25}\right) \left(\frac{0.1}{0.9}\right) = 3/9 = 1/3. \tag{15}$$

The log odds are $-\log 3$ and the Bayes decision is -1. Both worlds generate the same observed answer–reliability multiset with positive probability. A lineage-blind aggregator must return the same output on that multiset and therefore cannot be Bayes-optimal in both, proving Theorem 2.

B.3 Replica-Quotient Characterization of Replication Invariance

For the forward direction of Theorem 3, assume $\mathcal{A} = \bar{\mathcal{A}} \circ q_{\approx}$. If $T' \approx T$, adding T' does not create a new represented class:

$$q_{\approx}(S \cup \{T, T'\}) = q_{\approx}(S \cup \{T\}). \tag{16}$$

Applying $\bar{\mathcal{A}}$ proves invariance of both prediction and confidence.

Conversely, assume replication invariance. For any finite set of represented replica classes C, choose a multiset S_C containing one representative of every class and define $\bar{\mathcal{A}}(C) = \mathcal{A}(S_C)$. Any other multiset with quotient C differs only by adding or removing members of already represented replica classes. Repeated application of replication invariance therefore gives the same prediction and confidence, so $\bar{\mathcal{A}}$ is well defined and $\mathcal{A} = \bar{\mathcal{A}} \circ q_{\approx}$.

For additive instance counting, let the binary decision be $\operatorname{sign}(d)$ with $d = \sum_{U \in S} s(U)$. If a trace T has $s(T)$ with sign opposite to d, then adding any

$$k > \frac{|d|}{|s(T)|} \tag{17}$$

replicas flips the decision. Hence positive per-instance additive mass is replication manipulable unless contributions are first collapsed at the equivalence-class level.

B.4 Acquisition–Handling Decomposition

For any predictive distribution $q_{\mathcal{A}}$ based on (S, T_Q), the conditional cross-entropy identity gives

$$\begin{aligned} \mathcal{R}(q_{\mathcal{A}}, Q) = H(Y \mid S, T_Q) \\ + \mathbb{E}[D_{\mathrm{KL}}(p_Q \| q_{\mathcal{A}})] \,. \end{aligned} \tag{18}$$

Because $T_V = (T_Q, T_{V \setminus Q})$, the entropy difference is

$$H(Y \mid S, T_Q) - H(Y \mid S, T_V) = I(Y; T_{V \setminus Q} \mid S, T_Q). \tag{19}$$

Subtracting $\mathcal{R}^\star(V) = H(Y \mid S, T_V)$ proves Theorem 4. If a message T is generated through the Markov chain $Y \to (S, T_Q) \to T$, then $Y \perp\!\!\!\perp T \mid S, T_Q$ and its conditional mutual information is zero, establishing the closed-transcript limitation stated in the main text.

B.5 Conditional Information and Greedy Source Acquisition

Let V index candidate traces and define $F : 2^V \to \mathbb{R}_{\geq 0}$ by

$$F(Q) = I(Y; T_Q \mid S). \tag{20}$$

Normalization follows immediately: $F(\varnothing) = 0$. For any $A \subseteq V$ and $j \notin A$, the chain rule gives

$$\begin{aligned} F(A \cup \{j\}) - F(A) &= I(Y; T_j \mid S, T_A) \\ &= H(T_j \mid S, T_A) \\ &\quad - H(T_j \mid Y, S, T_A). \end{aligned} \tag{21}$$

Conditional mutual information is non-negative, so F is monotone without an independence assumption. Now assume that candidate traces are conditionally independent given (Y, S). Then

$$H(T_j \mid Y, S, T_A) = H(T_j \mid Y, S), \tag{22}$$

which is independent of A. If $A \subseteq B$, conditioning reduces entropy:

$$H(T_j \mid S, T_A) \geq H(T_j \mid S, T_B). \tag{23}$$

Therefore the marginal gain of j cannot increase as the selected set grows, establishing submodularity; in the terminology of Chen & Waggoner (2016), these signals are informational substitutes for log loss. The classical result of Nemhauser et al. (1978) then gives

$$F(Q_{\text{greedy}}) \geq (1 - 1/e) \max_{|Q| \leq k} F(Q) \tag{24}$$

for a cardinality budget k.

The guarantee holds for fixed-cardinality, non-adaptive selection over provenance-grouped, conditionally independent sources. It supplies a normative objective for ranking candidate lineages in future open-system acquisition policies; RIDE-CAL uses the frozen reliability- and uncertainty-based policy described in Appendix C.1.

C RIDE-CAL IMPLEMENTATION DETAILS

C.1 FROZEN SELECTOR AND ACTIVE ACQUISITION

The candidate model is an L_2-regularized logistic regression with $C = 0.1$. Its twelve standardized features are source identity (lexical, dense, or PRF), mean reported confidence, within-lineage agreement, support-verifier score, retrieval top score and margin, document count, mean input tokens, cross-lineage answer count, and summed document overlap. Candidate probabilities receive a Platt map before Equation 7; the winning score share receives a second task-level Platt map. Hyperparameters C, τ, λ, and both Platt maps are selected by nested cross-validation on 99 development tasks, using five outer and four inner folds over $C \in \{0.01, 0.1, 1, 10\}$, $\tau \in \{0.5, 1, 2\}$, and $\lambda \in \{0, 0.5, 1, 2\}$. The frozen full-panel selector has $\tau = 2$ and $\lambda = 2$.

Active RIDE-CAL processes lexical, complementary-PRF, and dense sources in that order. It continues when maximum candidate reliability is below 0.55, lineage-answer entropy exceeds 0.25, the top-two answer margin is below 0.25, or calibrated combined confidence is below 0.65. Development selection constrains mean acquired sources to at most two. The frozen policy uses 1.86 sources on confirmation.

The first algorithm specifies how the three retrieval lineages and their evidence fingerprints are constructed before inference.

Algorithm 1 Frozen retrieval-source construction

Require: Question q, candidate documents $\mathcal{D}$, normalized dense encoder e, $k = 4$, PRF terms $m = 8$, prompt budget $B = 6000$
Ensure: Lexical, complementary-PRF, and dense sources with evidence fingerprints
Score $\mathcal{D}$ by BM25($k_1 = 1.2, b = 0.75$) on q; let L be the top-k documents
Let d_1 be the top BM25 document; select the m highest-IDF non-stopwords from its title and text
Form expanded query $q' = q \,\|\, \text{title}(d_1) \,\|\, \text{terms}(d_1)$ and rank all documents by BM25 on q'
Let P be the top-k PRF ranking
if IDs(P) = IDs(L) and an outside PRF candidate exists **then**
    Replace the lowest-scoring member of P by the highest-ranked outside candidate
Encode q and every title–document text with e; let G be the top-k cosine ranking
for $X \in \{L, P, G\}$ **do**
    Append ranked sentences until the prompt reaches B tokens; truncate only the final sentence if needed
    $f_X \leftarrow \text{SHA256}(\text{document IDs}, \text{sentence IDs}, \text{truncation chars})$
return $(L, f_L), (P, f_P), (G, f_G)$ as lexical, complementary-PRF, and dense sources

The second algorithm then shows how fingerprints, replica removal, within-lineage disagreement, and cross-lineage overlap enter the frozen decision.

Algorithm 2 Hierarchical RIDE-CAL inference

Require: Task, trace multiset S, source metadata, frozen selector $\theta = (\beta, g_c, g_t, \tau, \lambda)$
Ensure: Normalized answer $\hat{a}$, calibrated confidence $\hat{p}$

```
Remove repeated records with an already represented trace identifier
Group remaining records and sources by rendered-evidence fingerprint ℓ
for all lineages ℓ do
    Group traces by normalized answer a and compute the frozen 12-dimensional candidate feature vector x_ℓa
    r_ℓa ← g_c(σ(β_0 + β^⊤ standardize(x_ℓa)))
    π_ℓ(a) ← softmax_a(logit(r_ℓa)/τ) and w_ℓ ← max_a r_ℓa
    o_ℓ ← Σ_{m≠ℓ} Jaccard(documents_ℓ, documents_m)
    for all candidate answers a in ℓ do
        s(a) ← s(a) + w_ℓ π_ℓ(a)/(1 + λ o_ℓ)
Choose â by maximum s(a); break ties by maximum candidate reliability, source priority lexical ≺ dense ≺ complementary-PRF, then answer string
p_raw ← s(â)/Σ_a s(a) and p̂ ← g_t(p_raw)
return (â, p̂), used trace identifiers, effective lineage count, and answer scores
```

C.2 FULL FACTORIAL RESULTS

For completeness, Table 4 expands the main-text factorial summary to every evidence cell and reports the accompanying token and NLL measures.

Table 4: Complete 2×2 evidence cells. Accuracy and token F1 are percentages. All four cells use the same solver settings within each task.

Dataset	Cell	n	Acc.	Token F1	Binary NLL	Input tok.
MuSiQue	R	159	7.55	15.89	3.553	329
	$R + D$	159	4.40	14.71	2.797	2,676
	$R + P^\star$	159	47.80	60.44	1.889	492
	$R + D + P^\star$	159	47.17	60.06	1.855	2,839
HotpotQA	R	100	16.00	30.72	4.010	203
	$R + D$	100	15.00	31.76	4.094	2,154
	$R + P^\star$	100	59.00	77.87	3.281	265
	$R + D + P^\star$	100	49.00	68.65	2.924	2,216

C.3 FULL NON-ORACLE CONFIRMATION RESULTS

Table 5 provides the complete frozen confirmation comparison, including individual retrieval baselines and all selector variants omitted from the compact main-text table.

Table 5: All frozen confirmation baselines over 200 tasks. Conflict accuracy uses the 115 tasks on which at least two source-modal answers disagree.

Method	Acc.	Conflict acc.	NLL	Brier	ECE-10
Lexical RAG	44.5	29.6	3.670	0.351	0.390
Dense RAG	39.0	21.7	3.781	0.371	0.431
Complementary-PRF RAG	43.5	26.1	3.390	0.349	0.381
Instance majority	50.5	39.1	1.646	0.252	0.187
Confidence-weighted instances	50.5	39.1	1.646	0.251	0.216
Exact-replica collapse	50.5	39.1	1.646	0.252	0.187
Hard lineage vote	46.0	30.4	2.321	0.278	0.274
Reliability-weighted lineage vote	51.0	40.0	3.330	0.380	0.387
Hard lineage selector	51.0	40.0	0.703	0.250	0.120
Verifier selector	40.0	25.2	1.615	0.349	0.341
RIDE-CAL	52.0	40.9	0.650	0.228	0.089
Active RIDE-CAL	53.0	42.6	0.662	0.234	0.081
Oracle best available	63.5	59.1	0.292	0.106	0.188

Across the 200 tasks, 60 six-trace panels are unanimous and 17 are unanimously wrong. Wrong consensus is therefore 8.5% unconditionally and 28.33% conditional on unanimity. Confidence weighting uses reported probabilities as vote weights; if all weights are zero, it falls back to unweighted voting (two of 600 source panels and no full panels). Ties follow the frozen source order, sample order, and normalized answer string.

C.4 Provenance Perturbation Endpoints

Table 6 lists the endpoint values underlying the robustness curves and anchors each perturbation label to its measured accuracy and NLL.

Table 6: Frozen provenance-perturbation endpoints. Exact-replication change is 0% for RIDE-CAL in every row.

Condition	Endpoint	Acc.	Binary NLL
Clean lineage labels	–	52.0	0.650
False merge	100% merged	48.5	0.636
Missing document provenance	100% removed	48.5	0.636
Ordinary false split	100% split	53.0	0.657
Adversarial split	100% split	50.5	0.658
Overlap graph	threshold 0.50	46.5	0.630
Overlap graph	threshold 0.75	52.0	0.650

False merges randomly combine source fingerprints at the stated rate; ordinary splits assign selected traces to new sublabels without using their answers; adversarial splits use answer disagreement to isolate competing traces. Missing provenance removes document identifiers before fingerprint construction. Partial-overlap grouping connects sources whose document-set Jaccard overlap exceeds the threshold and uses graph components as lineages. Each corruption level averages over 100 fixed random seeds. The result files include the full curves and frozen manifests recording the tasks, traces, sources, support scores, selector, and protocol.

D Controlled Study Details

D.1 Complete Controlled MuSiQue Results

Table 7 expands the controlled mechanism comparison to every solver and handling condition on MuSiQue.

Table 7: Complete controlled MuSiQue results over 159 questions. The designated-source rule matches Independent-only by construction; its contrast with no-lineage voting is a handling diagnostic.

Method	Acc.	F1	Wrong cons.	NLL	Calls	Input tok.
Single/shared	3.77	14.87	–	4.77	1	2,616
Same-model/shared	5.66	15.68	40.25	5.99	3	7,833
Heterogeneous/shared	3.77	14.79	11.32	2.15	3	7,950
Debate/shared	6.29	17.38	61.64	8.79	6	15,837
Qwen/shared	4.40	14.71	–	2.80	1	2,676
Independent-only	56.60	69.27	–	1.31	1	488
RIDE-CONTROLLED no-lineage	8.18	18.17	6.29	1.65	4	8,322
RIDE-CONTROLLED/designated	56.60	69.27	8.18	1.31	4	8,322

D.2 Cross-Dataset Extension

The frozen extension uses 100 HotpotQA and 100 GSM8K questions; all 1,400 records pass evidence-ID, raw-file, and token-cap audits. Table 8 reports the paired effects that connect these extensions to the controlled mechanism claims.

Table 8: Supplemental paired effects with 95% paired-bootstrap intervals.

Dataset	Effect	Δ pp	95% CI
HotpotQA	Acquisition	39.0	[29.0, 49.0]
HotpotQA	Controlled oracle contrast	30.0	[20.0, 40.0]
HotpotQA	Handling	28.0	[19.0, 38.0]
GSM8K	Tool source	40.0	[30.0, 50.0]
GSM8K	Handling	33.0	[24.0, 42.0]

HotpotQA bridge effects are 42.2, 33.3, and 31.1 points; ten comparison questions are inconclusive. GSM8K heterogeneous/shared, calculator, no-lineage, and RIDE-CONTROLLED reach 60%, 100%, 67%, and 100%. Replication changes instance voting in 58.5% of 1,600 attacks and RIDE-CONTROLLED in 0%.

Table 9: Complete HotpotQA results over 100 questions. Accuracy, token F1, and wrong consensus are percentages.

HotpotQA method	Acc.	F1	Wrong cons.	NLL	Calls
Single/shared	24.0	39.9	–	5.06	1
Same-model/shared	20.0	38.1	37.0	5.43	3
Heterogeneous/shared	24.0	40.6	22.0	3.54	3
Debate/shared	23.0	39.2	50.0	7.17	6
Qwen/shared	15.0	31.8	–	4.09	1
Independent-only	54.0	74.3	–	3.80	1
RIDE-CONTROLLED no-lineage	26.0	44.1	18.0	3.04	4
RIDE-CONTROLLED/designated	54.0	74.3	21.0	3.80	4

Table 10: Complete GSM8K results over 100 questions. Accuracy and wrong consensus are percentages; exact-match score equals accuracy.

GSM8K method	Acc.	Wrong cons.	NLL	API calls	Tool uses
Single/shared	59.0	–	4.97	1	0
Same-model/shared	56.0	4.0	0.85	3	0
Heterogeneous/shared	60.0	0.0	0.51	3	0
Independent tool	100.0	–	0.00	0	1
RIDE-CONTROLLED no-lineage	67.0	0.0	0.34	3	1
RIDE-CONTROLLED/tool	100.0	0.0	0.00	3	1

D.3 ILLUSTRATIVE EVIDENCE INTERVENTION

The following example grounds the abstract lineage distinction in one concrete question before the final intervention table.

For HotpotQA question *"Live Wire Radio will possibly be considered as a replacement for a variety show created in what year?"*, the gold answer is 1974. The shared context states that Live Wire Radio launched in 2004 and may replace *A Prairie Home Companion*, but withholds the sentence stating that *A Prairie Home Companion* was created in 1974. This example is illustrative rather than an additional statistical result.

Table 11: A concrete HotpotQA intervention showing how response multiplicity can erase newly acquired evidence when lineage is ignored.

Component	Observed evidence or answers	Decision
Shared lineage	Three DeepSeek traces answer 2004, "not specified," and 2004 from the incomplete context.	Shared majority: 2004
Independent lineage	Qwen observes the withheld creation-year sentence and answers 1974.	Independent source: 1974
Instance voting	The same four answer instances are counted separately.	2004 (wrong)
RIDE-CONTROLLED	The three incomplete-context traces form one lineage; the frozen source rule preserves the conflicting independent lineage.	1974 (correct)

Trace audit. The 1,831-record MuSiQue cache has no missing responses or token-cap violations; seven excluded-task records contribute zero rows. Audit verifies all 159 conditions, 1,272 shared and 159 independent traces, matched RIDE-CONTROLLED/no-lineage inputs, and all 900 HotpotQA and 500 GSM8K traces.