WHEN REPETITION BECOMES EVIDENCE: SEMANTIC EVIDENCE OVERCOUNTING IN MULTI-AGENT RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent retrieval-augmented generation (RAG) systems may treat agreement as independent evidence even when reports share a source. We study this behavioral-level mechanism in MADAM-RAG: when content derived from a single wrong semantic unit is copied or paraphrased across multiple surface-level documents, it produces apparently independent agent reports that may be counted repeatedly. We distinguish independent wrong semantic units K_w from surface-level wrong documents M_w in a TriviaQA-based environment with controlled provenance. On Qwen2.5-7B-AWQ, under No Debate and with $K_w = 1$, increasing M_w from 1 to 8 reduces accuracy from 90% to 47% and raises target-wrong selection from 0% to 48%; holding $M_w = 6$ and varying K_w yields no comparable monotonic trend. Conditional mutual information and Bayesian accumulation provide normative references, while a correlation-neglect model captures residual behavioral weight on copies. Matched Context and round-level states indicate that debate propagates or stabilizes an existing wrong-evidence advantage rather than inherently creating errors. Aggregation- and communication-stage interventions further support repeated exposure as part of the pathway to erroneous consensus. Together, these results support a predictive and falsifiable behavioral-level account of semantic evidence overcounting, without making claims about its internal neural implementation.

1 INTRODUCTION

Retrieval-augmented generation (RAG) supplements parametric knowledge with external documents but also creates multi-document evidence-integration problems (Lewis et al., 2020; Karpukhin et al., 2020; Izacard and Grave, 2021; Borgeaud et al., 2022; Asai et al., 2024; Shi et al., 2024). Position, noise, document count, and conflicts in long contexts can alter answers (Liu et al., 2024; Hsieh et al., 2024; Yoran et al., 2024; Cuconasu et al., 2024; Levy et al., 2025). More supporting documents therefore need not provide more independent evidence. What remains unclear is whether multi-agent RAG systems distinguish repeated reports of the same underlying evidence from genuinely independent evidence, and how a failure to make this distinction affects collective decisions.

This distinction is sharper in multi-agent RAG. MADAM-RAG assigns documents to agents and forms an answer through interaction and aggregation (Wang et al., 2025). Copies or paraphrases derived from one wrong semantic unit can therefore become multiple apparently independent reports (Figure 1). An aggregator that counts agents without accounting for provenance may mistake a repeated rationale for independent evidence.

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.

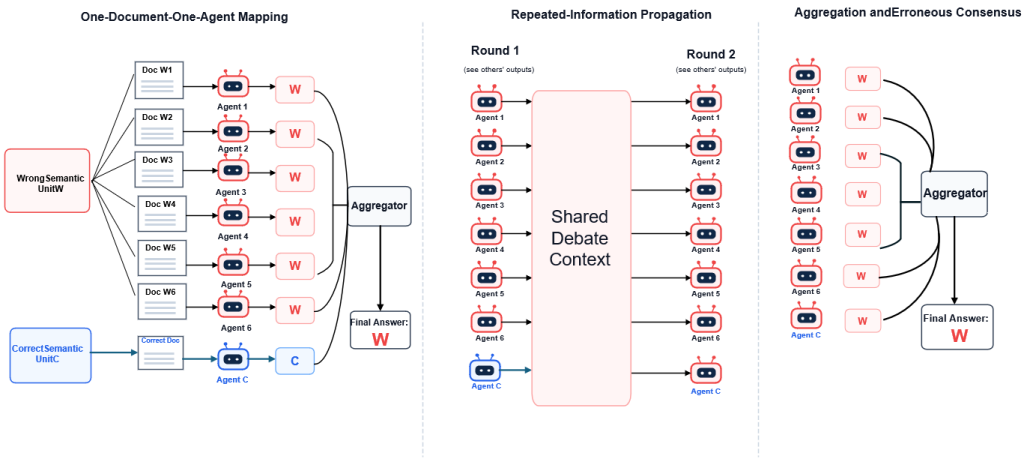


Figure 1: Schematic mechanism overview of erroneous consensus caused by surface duplication in multi-agent RAG; agent counts are simplified for illustration. A wrong semantic unit is copied or paraphrased into multiple documents, creating repeated reports across agents. In No Debate, repeated wrong reports can already bias Round-0 aggregation toward W. In Full Debate, repeated wrong semantics propagate through the shared debate context and can cause the initially correct agent to switch from blue to red. Repeated wrong reports may then be overcounted as independent evidence, leading to erroneous consensus. Red agents carry wrong information, whereas blue agents carry correct information.

We denote the independent semantic units supporting the target-wrong answer W by K_w and the surface-level wrong documents carrying them by M_w. When $M_w > K_w$, additional documents mainly increase repeated exposure. If shared provenance is not fully recognized, repeated reports retain residual behavioral weight that communication may propagate or stabilize.

$$K_w \leq M_w, \quad R_W = M_w - K_w \tag{1}$$

We identify the surface duplication effect, model incomplete correlation discounting, trace its propagation through communication, and test it with interventions at aggregation and communication. Manipulating K_w and M_w separates surface duplication from independent semantic variation. Conditional mutual information and dependence-aware Bayesian accumulation provide normative references, while ρ yields testable behavioral predictions. Matched Context characterizes communication effects; semantic grouping and novelty blocking test stage-specific dependence.

Our behavioral-level contributions are:

(i) separating surface-level wrong documents M_w from independent wrong semantic units K_w;

(ii) formalizing incomplete correlation discounting in a parsimonious behavioral model that predicts how repeated exposure can shift conditional C/W decisions despite low marginal information;

(iii) testing this account through input manipulations, communication controls, and two-stage interventions. ρ is a residual behavioral weight, not a neural parameter, and the interventions test the proposed mechanism rather than introduce new RAG methods.

2 Related Work

2.1 RAG and Multi-Document Evidence Integration

RAG, dense retrieval, Fusion-in-Decoder, RETRO, REPLUG, and Self-RAG extend external-memory retrieval and integration (Lewis et al., 2020; Karpukhin et al., 2020; Izacard and Grave, 2021; Borgeaud et al., 2022; Shi et al., 2024; Asai et al., 2024), while surveys emphasize relevance,

retrieval timing, and attribution (Fan et al., 2024). We instead study when surface document count misrepresents independent information.

2.2 Conflicting Evidence and the Persuasive Effect of Repetition

Under knowledge conflict, models trade off parametric knowledge against contextual evidence and other textual cues (Qian et al., 2023; Xie et al., 2024; Du et al., 2024; Xu et al., 2024; Su et al., 2024), and evidence frequency, wording, and salience affect perceived credibility (Jin et al., 2024; Wan et al., 2024; Naphade, 2026). MADAM-RAG addresses conflicting evidence (Wang et al., 2025); we use it to distinguish copies derived from one wrong semantic unit from independent wrong semantic units.

2.3 Multi-Agent Debate and Error Propagation

Multi-agent debate may improve reasoning or factuality (Chan et al., 2023; Du et al., 2024), but its gains depend on protocol, topology, and initial answers and are not consistently better than simple ensembles (Smit et al., 2024; Li et al., 2024). Adversarial agents, conformity, identity cues, and consensus rules also affect group outputs (Amayuelas et al., 2024; Weng et al., 2025; Choi et al., 2026; Cui et al., 2026).

We track how repeated information changes wrong-answer adoption and retention across rounds.

2.4 Research Gap

Prior work separately studies conflicting documents, evidence frequency, and agent interaction, but not the full pathway from a shared wrong semantic unit to multiple agent reports, residual weight caused by incomplete dependence discounting, and subsequent propagation through communication and aggregation. Separating K_w from M_w enables measurement; conditional mutual information and correlation neglect provide theory; state transitions and two-stage interventions test the mechanism.

Unlike deduplication or reranking work evaluated mainly by final performance, we ask why an undeduplicated system fails and use provenance labels for identification. Grouping and gating test whether the explanation is manipulable; they do not claim algorithmic superiority.

3 Controlled Study Design and Theoretical Mechanism

3.1 System, Task, and Notation

In standard MADAM-RAG, J=10 document-conditioned agents each read one document and produce a Round 0 answer; debate adds two message-exchange rounds before final aggregation. Main experiments use Qwen2.5-7B-AWQ with temperature=0 (Qwen Team, 2025). Outputs are C, W, or OTHER, with correct evidence fixed at $M_c = 2$ and $K_c = 1$.

We hold J, M_c, K_c, and debate rounds fixed while varying M_w or K_w, filling remaining slots with neutral documents. The duplication branch uses $K_w = 1$ and $M_w \in \{1, 2, 4, 6, 8\}$; the independent-semantic control fixes $M_w = 6$ and uses $K_w \in \{1, 2, 3, 4\}$. Each condition has 100 questions. The two branches manipulate, respectively, the number of repeated carriers of one wrong semantic unit and the number of independent wrong semantic units under a fixed wrong-document budget.

3.2 TriviaQA-Based Evidence with Controlled Provenance

We select unambiguous TriviaQA questions for which target-wrong evidence can be constructed (Joshi et al., 2017). In our controlled generator, copies and paraphrases derived from the same wrong semantic unit share a semantic_unit_id, which serves as a provenance label; distinct semantic units receive different IDs. IDs are hidden from ordinary agents and the baseline Aggregator and used only to manipulate K_w, compute mechanism metrics, and implement oracle interventions.

This design provides provenance ground truth rather than assuming that real systems know K_w. Independent variation of K_w and M_w separates new semantics from repeated exposure; in real

retrieval, latent K_w must be estimated from metadata, reposting chains, timestamps, or semantic clustering.

At fixed $K_w = 1$, changing M_w alters only repeated exposure to the same wrong semantic unit; at fixed $M_w = 6$, changing K_w reallocates six documents across semantic units. Agent count, correct evidence, questions, and decoding remain fixed, ruling out changes in agent count as an explanation for the observed repetition effect.

C/W/OTHER record the selected answer, whereas semantic_unit_id records the provenance of text supporting W. Both are needed to distinguish six independent rationales from six copies and to trace overcounting from documents through agent states to aggregation.

3.3 Normative References: Conditional Mutual Information and Dependence-Aware Bayesian Accumulation

For the $K_w = 1$ duplication branch, let $Y \in \{C, W\}$, $D_1, \ldots, D_{M_w}$ support W, and U denote their shared semantic unit. After the first document generated from source U, subsequent same-source copies should add almost no additional information about the answer:

$$I(Y; D_j \mid D_1, \ldots, D_{j-1}) \approx 0, \quad j \geq 2 \tag{2}$$

Equation (2) is a normative constraint for our controlled generator, not a claim that every paraphrase is uninformative. The mutual-information chain rule gives:

$$I(Y; D_{1:M_w}) = \sum_{j=1}^{M_w} I(Y; D_j \mid D_{1:j-1}) \tag{3}$$

Under Equation (3), with $K_w = 1$, the first copy carries most of the information; later same-source copies add little. Thus, surface document count may grow without comparable independent information. Let L_0 denote the baseline W/C log-odds from the prior and fixed correct evidence. A provenance-aware Bayesian update uses the joint likelihood:

$$L_W = \log \frac{P(W \mid D_{1:M_w})}{P(C \mid D_{1:M_w})} = L_0 + \log \frac{p(D_{1:M_w} \mid W, U)}{p(D_{1:M_w} \mid C, U)} \tag{4}$$

Treating dependent documents as independent instead yields $q(D_{1:M_w}|Y) = \prod_{j=1}^{M_w} q(D_j|Y)$,

repeatedly counting one source (Cover & Thomas, 2006; Pearl, 1988; Koller &

Friedman, 2009):

$$L_{\text{naive}} = L_0 + \sum_{j=1}^{M_w} \log \frac{q(D_j \mid W)}{q(D_j \mid C)} \tag{5}$$

Equations (4) and (5) contrast dependence-aware and independent accumulation. We do not claim that the model performs probabilistic updating; we test whether its behavior shows analogous correlation neglect, such that same-source copies retain influence beyond their incremental information.

3.4 A Behavioral Model of Correlation Neglect

Let $\rho \in [0, 1]$ be the average behavioral weight retained by same-source copies. If an independent wrong semantic unit k appears n_k times, with $\sum_k n_k = M_w$, and has strength $\ell_{W,k}$,the generalized model is:

$$L_W = L_0 + \sum_{k=1}^{K_w} [1 + \rho(n_k - 1)] \ell_{W,k} \tag{6}$$

Equation (6) allows semantic units to differ in persuasiveness. At $K_w = 1$, the first instance contributes ℓ_W and each of the remaining $M_w - 1$ copies retains fraction ρ, giving:

$$L_W(M_w) = L_0 + [1 + \rho(M_w - 1)]\ell_W, \quad 0 \leq \rho \leq 1 \tag{7}$$

ρ=0 is complete discounting, ρ=1 treats copies as independent, and $0 < \rho < 1$ is partial discounting. Equation (7) implies the marginal wrong-answer advantage:

$$\frac{\partial L_W}{\partial M_w} = \rho \ell_W \tag{8}$$

Under a binary approximation that compares only C and W, conditional accuracy and its directional prediction are:

$$P(C \mid Y \in \{C, W\}, M_w) = \frac{1}{1 + \exp[L_w(M_w)]} \tag{9}$$

$$\frac{\partial P(C \mid Y \in \{C, W\}, M_w)}{\partial M_w} = -\frac{\ell'_w \exp[L_w(M_w)]}{(1 + \exp[L_w(M_w)])^2} < 0 \tag{10}$$

Thus, if $\rho > 0$ and $\ell_W > 0$, additional copies increase W's relative advantage and reduce conditional accuracy. ρ links theory to observed residual behavioral weight without specifying an internal reasoning process.

3.5 Communication Stage: State-Dependent Modulation of Repeated Information

Let $G_T(M_w, Z_0)$ denote how T communication rounds modulate an existing wrong-answer advantage given initial agent state Z_0 and repetition M_w:

$$L_W^{(T)} = L_0 + G_T(M_w, Z_0)[1 + \rho(M_w - 1)]\ell_W \tag{11}$$

G_T models state-dependent communication effects without fixing their direction. We test whether genuine peer content changes wrong-answer adoption and retention relative to Matched Context with matched rounds, messages, and approximately matched tokens.

3.6 Falsifiable Predictions

The model yields three falsifiable predictions. P1: at fixed K_w, increasing M_w should increase W's relative advantage if copies are incompletely discounted. P2: if communication propagates that advantage, Full Debate and Matched Context should differ after matching rounds, messages, and tokens. P3:if repetition operates across stages, reducing report multiplicity or repeated-message exposure should reduce the wrong-answer advantage; semantic-aware interventions should further outperform matched random controls if dependency information itself matters

3.7 Mapping Theoretical Quantities to Observations

The empirical implications of the latent quantities L_W, ρ, and G_T are tested indirectly through outputs, agent transitions, and interventions.

Holding K_w fixed while varying M_w tests repetition sensitivity; regression isolates the repetition term; and state changes plus Matched Context characterize communication-mediated propagation.

4 PHENOMENON AND MECHANISTIC EVIDENCE

4.1 SURFACE DUPLICATION EFFECT AND INDEPENDENT-SEMANTIC CONTROL

At fixed $K_w = 1$, increasing $M_w \in \{1, 2, 4, 6, 8\}$ steadily lowers accuracy and raises target-wrong selection. Because all wrong documents share semantic_unit_id, the manipulation adds repeated exposure but no independent wrong information.

At fixed $M_w = 6$, varying $K_w \in \{1, 2, 3, 4\}$ does not produce the same monotonic decline (Figure 2(b)); M_w and K_w therefore have different behavioral effects.

Under duplication, lost accuracy mainly becomes W rather than OTHER (Figure 2(a)), consistent with added weight on repeated wrong semantics rather than diffuse uncertainty.

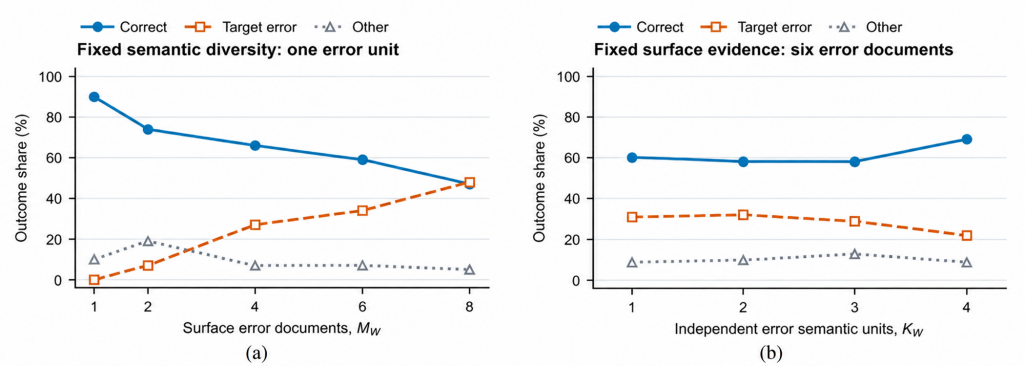


Figure 2: Surface-level wrong-document count and independent wrong-semantic count have different effects on final outputs.

4.2 BEHAVIORAL REGRESSION AS A DIAGNOSTIC OF RESIDUAL REPETITION WEIGHT

After controlling for independent semantic units K_w, we test whether repeated exposure $M_w - K_w$ predicts binary C/W outcomes after excluding OTHER:

$$\text{logit}\, P(W) = \beta_0 + \gamma_{\text{ind}} K_w + \gamma_{\text{rep}} (M_w - K_w) \tag{12}$$

We obtain $\gamma_{\text{ind}} = 0.272$ and $\gamma_{\text{rep}} = 0.432$. Since $\gamma_{\text{rep}} > 0$, repeated exposure remains positively associated with selecting W after controlling for independent semantics, consistent with residual influence from same-source copies.

This regression diagnoses behavioral repetition sensitivity, not internal computation. Equation (12) empirically separates independent-semantic and repeated-exposure contributions, complementing the idealized accumulation in Equation (7).

4.3 INCREMENTAL EFFECT OF GENUINE PEER CONTENT

We compare No Debate, which aggregates Round 0; Matched Context, which matches rounds, messages, and tokens while replacing peer messages; and Full Debate, which uses genuine interactions. Accuracy declines in all three conditions as M_w increases; under high repetition, Full Debate is lower than Matched Context (Table 1), suggesting an additional effect of genuine peer content.

4.4 ROUND-BY-ROUND STATES: ADOPTION, RETENTION, AND REPEATED CARRIERS

Let $Z_i^{(t)}$ be agent is state at round t. State transitions define wrong-answer adoption a_t, retention b_t, and repeated wrong carriers R_t:

Table 1: Final-answer accuracy under different communication conditions.

M_w	**No Debate**	**Matched Context**	**Full Debate**
1	90%	82%	87%
2	74%	76%	71%
4	66%	65%	61%
6	59%	54%	49%
8	47%	50%	45%

$$a_t = \frac{N(Z^{t-1} \neq W, Z^t = W)}{N(Z^{t-1} \neq W)}, \quad b_t = \frac{N(Z^{t-1} = W, Z^t = W)}{N(Z^{t-1} = W)} \tag{13}$$

$$R_t = \max(N_W^t - K_w, 0) \tag{14}$$

Adoption is 4.55%–16.16% in the first round and 2.99%–25.35% in the second (Figure 3). Its non-monotonic dependence on M_w and round shows that peer effects depend on the state distribution and interaction stage rather than a fixed amplification process.

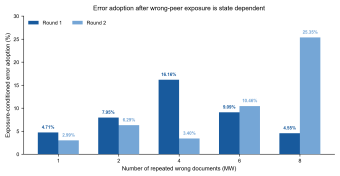


Figure 3: Wrong-answer adoption varies across repetition levels and communication rounds.

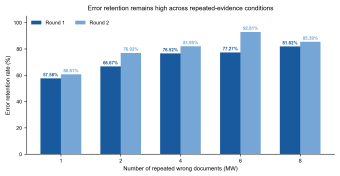


Figure 4: Existing erroneous states show high retention during communication.

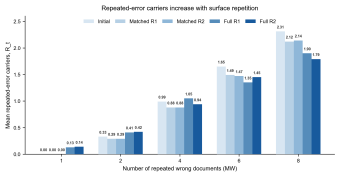


Figure 5: Excess W-supporting agents increase with surface repetition and persist after communication.

Let $b_c(M_w)$ denote the corrected W/C log-odds under communication condition c at repetition level M_w. We define the incremental peer-content effect and its contrast between two repetition levels as:

$$A_{\text{peer}}(M_w) = b_{\text{Full}}(M_w) - b_{\text{Matched}}(M_w), \quad DID_{\text{peer}}(m_a, m_b) = A_{\text{peer}}(m_a) - A_{\text{peer}}(m_b) \tag{15}$$

$A_{peer}(M_w)$ measures the incremental target-wrong advantage from genuine peer content at repetition level M_w, while $DID_{peer}(m_a, m_b)$ measures how this effect changes between two repetition levels.

Table 2: Corrected W/C log-odds and incremental peer-content effect.

M_w	$b_{\text{Matched}}(M_w)$	$b_{\text{Full}}(M_w)$	$A_{\text{peer}}(M_w)$
1	-3.160	-3.219	-0.059
2	-3.085	-1.249	$+1.835$
4	-1.441	-0.769	$+0.672$
6	-0.648	-0.251	$+0.397$
8	-0.441	-0.336	$+0.105$

As Table 2 shows, $A_{peer}(1)$=-0.059, providing little evidence of an additional shift toward the target-wrong answer from genuine peer content when evidence is nearly balanced.

$A_{peer}(M_w)$ is positive at all other repetition levels and peaks at M_w=2, where $A_{peer}(2)$=1.835, before declining as M_w increases. Thus, genuine peer content usually adds a target-wrong advantage, but not monotonically with repetition.

For the high-versus-low repetition contrast, DID_{peer}(8,1)=0.163, indicating a modest increase in the incremental influence of genuine peer content under high repetition. The appendix reports the calculation, interval estimates, and resampling procedure for DID_{peer}.

5 Mechanism Interventions Across Two Stages

5.1 Intervention Logic

We intervene at two distinct stages. At aggregation, we freeze the same Full Debate outputs and vary only grouping information. During communication, we keep Round 0 fixed and gate repeated semantics across the same two rounds. Stage-specific effects support a cross-stage pathway rather than a single-stage artifact.

5.2 Aggregation-Stage Intervention

After identical Full Debate, we freeze the 10 final reports and rerun only the Aggregator under Full Debate, Size-Matched Random, and True Semantic Grouping. Random preserves true group sizes but randomizes membership; True uses anonymized semantic_unit_id. No report is removed. Size-Matched Random accuracy is averaged over 20 randomizations; the other conditions are deterministic.

Table 3: Final-answer accuracy under different aggregation-grouping conditions.

M_w	**Full Debate**	**Size-Matched Random**	**True Semantic Grouping**	**True−Full**
1	87%	88.4%	89%	+2%
2	71%	84.25%	84%	+13%
4	61%	75.9%	76%	+15%
6	49%	69.2%	67%	+18%
8	45%	54.65%	54%	+9%

Random–True differences test whether aggregation is sensitive to the correctness of semantic dependency structure beyond group size alone.

5.3 Communication-Stage Intervention

All conditions share the same Round 0 and two communication rounds. Full Communication blocks nothing. Matched Random Gate matches deletion count and message features but ignores semantic identity. Its reported accuracy is averaged over 20 random masks. Perfect Novelty Gate blocks only repeated atomic arguments already seen by each receiver, using semantic_unit_id and the same masking template.

Table 4: Final-answer accuracy under communication-stage gating conditions.

M_w	**Full Debate**	**Random Gate**	**Novelty Gate**	**Novelty−Full**
1	87%	89%	90%	+3%
2	71%	83%	83%	+12%
4	61%	74%	75%	+14%
6	49%	70%	72%	+23%
8	45%	63.2%	64%	+19%

Novelty Gate improves final-answer accuracy over Full Debate at every M_w, with larger gains under high repetition (Table 4). Random Gate also improves performance, supporting exposure reduction but not superiority of provenance-aware gating over all non-semantic sparsification.

5.4 Closing the Mechanistic Loop

Together, the two interventions show that reducing effective report multiplicity at aggregation and reducing repeated-message exposure during communication can attenuate erroneous consensus. However, the matched random controls recover most of the same gains. Therefore, the intervention results support a cross-stage role for repetition and exposure, but do not by themselves establish that explicit semantic-provenance recognition is necessary at either stage.

6 Discussion and Conclusion

6.1 What the Mechanistic Explanation Supports

Results support a bounded behavioral account: multiple documents and agent reports derived from the same semantic unit retain residual behavioral weight under incomplete correlation discounting, and communication can propagate or stabilize this advantage. The input manipulation and regression diagnose sensitivity to repeated exposure, while the stage-specific interventions show that reducing effective multiplicity or repeated-message exposure attenuates erroneous consensus. Importantly, True Semantic Grouping and Perfect Novelty Gating are not consistently superior to their matched random controls; therefore, the intervention results support the role of repetition and exposure, but do not uniquely attribute the improvement to explicit provenance-aware processing.

6.2 Alternative Explanations and Mechanism Boundaries

Long context, position, token burden, wording, and content complexity may affect evidence use. Matched Context controls rounds, messages, and tokens but cannot eliminate all such effects. A majority heuristic explains sensitivity to report count, whereas incomplete correlation discounting explains why shared-source repetition becomes a distorted evidence proxy. The near-parity between semantic interventions and their matched random controls further suggests that generic count reduction or communication sparsification may explain part of the intervention gain. Thus, our evidence identifies repeated exposure as mechanistically relevant, but does not establish semantic provenance recognition as the unique causal ingredient of the intervention effect.

semantic_unit_id makes K_w controllable in our experiment, but real retrieval requires estimating provenance from metadata, temporal or propagation relations, embedding clusters, natural language inference (NLI), or provenance graphs. Oracle K_w is used for identification, not assumed available in deployment.

We validate the phenomenon on Qwen2.5-7B-AWQ, GPT-4o-mini, and Mistral-7B-Instruct-v0.3 (appendix). Because experiments use controlled TriviaQA, a fixed agent count, and limited communication rounds, they do not establish generality to all architectures or open-web retrieval.

6.3 Distinguishing Theoretical and Empirical Quantities

In Equation (7), ρ is a theoretical residual weight; the regression coefficient is an empirical association, not the same quantity. Equation (9) models conditional C/W competition rather than all output classes.

Direct estimation of ρ would require finer-grained probability observations, stronger identifiability, and a multiclass hierarchical model.

6.4 Conclusion

We separate surface-level wrong documents M_w from independent wrong semantic units K_w and show that same-source copies increase the target-wrong advantage in controlled MADAM-RAG. Normative references, a correlation-neglect model, communication controls, and two-stage interventions jointly support semantic evidence overcounting as an observable and manipulable behavioral-level mechanism. Its cross-model generality, provenance measurement in natural retrieval, and internal implementation remain open.

REFERENCES

Amayuelas, A. et al. MultiAgent Collaboration Attack. Findings of EMNLP, 2024.

Asai, A. et al. Self-RAG: Learning to Retrieve, Generate, and Critique Through Self-Reflection. ICLR, 2024.

Borgeaud, S. et al. Improving language models by retrieving from trillions of tokens. ICML, 2022.

Chan, C.-M. et al. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. arXiv:2308.07201, 2023.

Choi, H. K., Zhu, J. & Li, S. When Identity Skews Debate: Anonymization for Bias-Reduced Multi-Agent Reasoning. ACL, 2026.

Cover, T. M. & Thomas, J. A. Elements of Information Theory. Wiley, 2006.

Cuconasu, F. et al. The Power of Noise: Redefining Retrieval for RAG Systems. SIGIR, 2024.

Cui, Y. et al. Free-MAD: Consensus-Free Multi-Agent Debate. Findings of ACL, 2026.

Du, K. et al. Context versus Prior Knowledge in Language Models. ACL, 2024.

Du, Y. et al. Improving Factuality and Reasoning in Language Models Through Multiagent Debate. ICML, 2024.

Fan, W. et al. A Survey on RAG Meeting LLMs. KDD, 2024.

Hsieh, C.-Y. et al. Found in the Middle: Calibrating Positional Attention Bias Improves Long Context Utilization. Findings of ACL, 2024.

Izacard, G. & Grave, E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. EACL, 2021.

Jin, Z. et al. Tug-of-War between Knowledge: Knowledge Conflicts in Retrieval-Augmented Language Models. LREC-COLING, 2024.

Joshi, M. et al. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. ACL, 2017.

Karpukhin, V. et al. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP, 2020.

Koller, D. & Friedman, N. Probabilistic Graphical Models: Principles and Techniques. MIT Press, 2009.

Levy, S. et al. More Documents, Same Length: Isolating the Challenge of Multiple Documents in RAG. arXiv:2503.04388, 2025.

Lewis, P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS, 2020.

Li, Y. et al. Improving Multi-Agent Debate with Sparse Communication Topology. Findings of EMNLP, 2024.

Liu, N. F. et al. Lost in the Middle: How Language Models Use Long Contexts. TACL, 2024.

Naphade, A. Rational Synthesizers or Heuristic Followers? LLMs in RAG-based Question Answering. Findings of ACL, 2026.

Pearl, J. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, 1988.

Qian, C., Zhao, X. & Wu, S. T. 'Merge Conflicts!' Exploring the Impacts of External Distractors to Parametric Knowledge Graphs. arXiv:2309.08594, 2023.

Qwen Team. Qwen2.5 Technical Report. arXiv:2412.15115, 2025.

Shi, W. et al. REPLUG: Retrieval-Augmented Black-Box Language Models. NAACL, 2024.

Smit, A. P. et al. Should We Be Going MAD? A Look at Multi-Agent Debate Strategies for LLMs. ICML, 2024.

Su, Z. et al. ConflictBank: Evaluating the Influence of Knowledge Conflicts in LLMs. NeurIPS Datasets and Benchmarks, 2024.

Wan, A., Wallace, E. & Klein, D. What Evidence Do Language Models Find Convincing? ACL, 2024.

Wang, H. et al. Retrieval-Augmented Generation with Conflicting Evidence. COLM, 2025.

Weng, Z., Chen, G. & Wang, W. Do as We Do, Not as You Think: The Conformity of Large Language Models. ICLR, 2025.

Xie, J. et al. Adaptive Chameleon or Stubborn Sloth: Revealing LLM Behavior in Knowledge Conflicts. ICLR, 2024.

Xu, R. et al. Knowledge Conflicts for LLMs: A Survey. EMNLP, 2024.

Yoran, O. et al. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. ICLR, 2024.

AI ASSISTANCE

Generative AI tools were used during the preparation of this work to assist with language polishing and academic writing, literature retrieval and discovery, research ideation and experimental workflow development, drafting and revising portions of the manuscript, and generating controlled synthetic evidence variants used in the experiments. For example, AI tools were used to help identify potentially relevant work on retrieval-augmented generation, multi-agent debate, knowledge conflict, and repeated-evidence effects; to refine the presentation of the experimental workflow and descriptions of mechanisms such as surface duplication and communication-stage propagation; and to assist in drafting and revising selected parts of the manuscript.

For dataset construction, the base questions were drawn from TriviaQA and were not generated by AI. Generative AI was used only to assist in constructing controlled target-wrong evidence variants under predefined experimental constraints, including direct-copy and meaning-preserving paraphrase forms derived from the same underlying wrong semantic unit. In these variants, the underlying semantic evidence and intended target-wrong answer were kept fixed while the surface realization was varied, allowing different surface-level documents to share the same semantic_unit_id and enabling controlled separation between surface-level document count and the number of independent wrong semantic units. The generated evidence variants were subsequently reviewed by the authors to verify that they supported the intended target-wrong answer, preserved the required semantic relation, and satisfied the experimental controls.

All AI-assisted outputs, including writing suggestions, literature suggestions, experimental descriptions, and generated evidence variants, were subsequently reviewed, verified, and revised by the authors where necessary. References suggested through AI-assisted retrieval were independently checked against the original sources before inclusion. The authors made the final decisions regarding the research questions, methodology, experimental design, dataset labels and semantic-unit assignments, statistical analysis, interpretation of results, and conclusions, and take full responsibility for the content of this paper.

APPENDIX

This appendix documents the system configuration, controlled dataset construction, prompts and implementation details, mechanism interventions, and supplementary results. The main paper retains the core evidence chain from surface repetition to behavioral mechanism, communication-stage propagation, and stage-specific intervention. Here we provide model and decoding settings, output

parsing, the controlled construction of K_w and M_w, the definition of semantic_unit_id, a concrete data example, full prompts, intervention pseudocode and input transformations, and cross-model and peer-content analyses. Unless stated otherwise, all numerical values are taken from the archived formal runs.

APPENDIX ROADMAP

- Appendix A: system workflow, runtime configuration, model/tokenizer identity, and output normalization;
- Appendix B: controlled dataset, K_w, M_w, semantic_unit_id, surface-form statistics, and a concrete record;
- Appendix C: Experiment 1 agent and Aggregator prompts, communication-update prompt, and communication-stage Aggregator prompt;
- Appendix D: pseudocode for the four Experiment 3 interventions and paired input transformations;
- Appendix E: cross-model results and the incremental peer-content analysis for Full Debate versus Matched Context.

A SYSTEM WORKFLOW AND RUNTIME CONFIGURATION

A.1 SYSTEM WORKFLOW

Each question instantiates J = 10 document-conditioned agents, each with one private document. Agents answer independently at Round 0. Communication experiments add two rounds of peer-message exchange, after which the Central Aggregator produces the system answer from the ten final reports. Outputs are classified as the correct answer (C), the predefined target-wrong answer (W), or OTHER.

Table A1: Main experimental configuration.

Item	Experiment 1	Experiment 2	Experiment 3
Model	Qwen2.5-7B-AWQ	Qwen2.5-7B-AWQ	Qwen2.5-7B-AWQ
Inference interface	OpenAI-compatible; local vLLM	OpenAI-compatible	Reuses the formal model run
Temperature	0.0	0.0	Inherited from the source run
Maximum agent tokens	128	96	Inherited from the source run
Maximum Aggregator tokens	192	96	Inherited from the source run
Number of agents	10	10	10
Communication rounds	None	2	2 or final-aggregation intervention
Bootstrap replicates	2,000	2,000	2,000
Continuity correction	Not applicable	0.5	0.5

A.2 MODEL AND TOKENIZER IDENTITY

Experiments 2 and 3 use /root/autodl-tmp/MADAM-RAG/models/Qwen2.5-7B-AWQ. The Experiment 2 matching audit records the same tokenizer hash.

A.3 OUTPUT NORMALIZATION AND CLASSIFICATION

Experiment 1 first lowercases text, removes punctuation and the English articles a/an/the, and collapses whitespace. An output is classified directly when its normalized form exactly matches C or

W; an output containing only C or only W is also assigned to the corresponding class. Outputs that contain both or neither are assigned to OTHER.

Experiments 2 and 3 use a stricter final-label rule: the normalized output must exactly match C or W, otherwise it is OTHER. The JSON parser accepts answer/final_answer and reasoning/reason/explanation fields. If JSON parsing fails, an Answer line may be extracted, while parse_ok retains the failure state. All three formal pipelines preserve the raw output and never substitute a fabricated answer for a parse failure.

B Controlled Dataset and Source-Level Semantic Labels

B.1 Dataset Composition

The dataset contains 100 unique base questions and five M_w conditions $\{1, 2, 4, 6, 8\}$ per question, yielding 500 condition records. The base questions are drawn from TriviaQA.

Table B1: Dataset source and condition counts.

Source or condition	Unique base questions	Condition records
$M_w = 1$	100	100
$M_w = 2$	100	100
$M_w = 4$	100	100
$M_w = 6$	100	100
$M_w = 8$	100	100
Total	100	500

All conditions fix J = 10, M_c = 2,K_c = 1, and K_w = 1. The E1-A manipulation therefore changes only the number of surface carriers of the target-wrong semantic unit, M_w, and not the number of independent wrong semantic units.

B.2 K_w, M_w, and Semantic Units

Let D_w be the set of target-wrong documents and g(d) the underlying semantic-unit label of document d. Then

$$M_w = |D_w|, \qquad K_w = |\{g(d) : d \in D_w\}|.$$

semantic_unit_id is controller-side provenance ground truth and is not included in ordinary agent prompts. It separates the number of surface carriers from the number of independent semantic units. In a deployed system, this label would normally need to be estimated from provenance metadata, semantic clustering, or deduplication. True Semantic Grouping and Perfect Novelty Gate therefore test whether limiting repetition is effective when latent semantic identity is available; they do not show that provenance identification has already been solved in realistic retrieval.

B.3 Surface Variants and Text Statistics

paraphrase-v5 provides eight surface forms for each target-wrong semantic unit: V1/V3/V5/V7 are exact copies and V2/V4/V6/V8 are meaning-preserving paraphrases.

Table B2: Mean word count by document type.

Document type	Mean word count
Correct evidence	10.02
Target-wrong evidence	9.83
Neutral evidence	11.31

The three document classes are similar in mean length, but these statistics are not a human agreement assessment of semantic quality. The archive contains generation labels but no independent double-annotation record; the paper therefore does not claim measured inter-annotator agreement.

B.4 A Concrete Data Record

Table B3 shows record E1BQ0301__MW8 from E1-A. The question is "Who is the artist of the 'Freedom Suite'?"; C is David S. Ware and W is John Coltrane. The condition fixes J = 10, with M_c= 2 correct-evidence carriers,M_w = 8 target-wrong carriers, and no neutral documents. All eight wrong documents derive from the same underlying wrong semantic unit U1, so K_w remains 1 despite M_w = 8. The record stores question metadata, experimental variables, semantic provenance labels, candidate pools, and condition composition for controlled reconstruction.

Table B3: Data structure of one experimental condition record.

Field	Example Value	Meaning
`schema_version`	`Experiment1_E1A_paraphrase_v5_20260914`	Dataset schema/version identifier.
`record_type`	`E1A_paired_condition_record`	Experimental record type.
`base_qa_id`	`E1BQ0301`	Identifier of the base question.
`condition_record_id`	`E1BQ0301__MW8`	Identifier of this specific experimental condition.
`source_dataset`	`TriviaQA`	Source dataset of the base question.
`source_audit_id`	`TriviaQA_line_0024`	Source-level audit identifier.
`question`	Who is the artist of the "Freedom Suite"?	Question presented in this QA instance.
`C`	David S. Ware	Correct answer.
`W`	John Coltrane	Predefined target-wrong answer.
J	10	Total number of document-conditioned agents.
M_C	2	Number of correct-evidence document carriers.
K_C	1	Number of independent correct semantic units.
M_W	8	Number of target-wrong document carriers used in this condition.
K_W	1	Number of independent target-wrong semantic units.
M_N	0	Number of neutral documents used in this condition.
`wrong_semantic_units`	`U1`	Underlying wrong semantic unit from which the target-wrong variants are derived.
`condition_spec.composition`	`2C+8W+0N`	Condition-level document composition.
`condition_spec.wrong_unit_ids`	`[U1]`	Wrong semantic units activated in this condition.
`required_wrong_documents`	8	Number of wrong documents selected for this condition.
`required_neutral_documents`	0	Number of neutral documents selected for this condition.
`correct_document_pool`	2 document objects	Candidate correct-evidence documents.
`wrong_document_pool`	8 document objects	Candidate wrong-evidence surface variants derived from `U1`.
`neutral_document_pool`	7 candidate document objects	Shared candidate neutral-document pool; none are used when $M_N = 0$.

C Prompts and Implementation Details

```
You are one independent question-answering agent in Round 0.
Use only the assigned document below. Do not use outside knowledge and do not
infer credibility from writing style.
If the document does not contain information that answers the question, answer
UNKNOWN.
Return exactly one JSON object with keys "answer" and "reason". Keep the answer
short.

Question:
{question}

Assigned document:
{document}
```

Figure C1: Experiment 1 Round-0 agent prompt.

```
You are the Central Aggregator. Ten independent Round-0 agents each read
one document.
Resolve the question using only their answers and reasons below. Assess whether
each reason directly supports the relation asked in the question.
Do not use outside knowledge. Do not follow a simple majority rule mechanically,
and do not infer credibility from style or confidence.
Return exactly one JSON object with keys "answer" and "reason". Keep the answer
short.

Question:
{question}

Round-0 responses:
{responses}
```

Figure C2: Experiment 1 Central Aggregator prompt.

```
You are one QA agent in a multi-agent discussion.
Question: {question}
Your assigned document: {document}
Your previous response:
{own_previous}
Messages from the other nine agents, in the fixed sender order:
{peer_messages}
Reassess the answer using your document and the messages. Return exactly one JSON object:
{"answer":"one short answer or UNKNOWN","reasoning":"one concise sentence"}
Do not mention the experiment, document labels, or control condition.
```

Figure C3: Communication-update prompt.

```
You are the central answer aggregator.
Question: {question}
Final reports from ten agents:
{agent_messages}
Choose the single best answer using only these reports. Return exactly:
{"answer":"one short answer or UNKNOWN","reasoning":"one concise sentence"}
Do not mention voting counts, experimental conditions, or hidden document labels.
```

Figure C4: Communication-stage Central Aggregator prompt.

D INTERVENTION DETAILS

D.1 PSEUDOCODE FOR THE FOUR INTERVENTIONS

True Semantic Grouping and Size-Matched Random Grouping act only at final aggregation, whereas Perfect Novelty Gate and Matched Random Gate act during inter-agent communication. Algorithms D1–D4 preserve the stage-specific controls stated in the main paper.

Algorithm D1 True Semantic Grouping

Require: Frozen reports R = {r1, …, rJ}; controller-side semantic-unit map u(r)
Ensure: Prediction y_true from all J reports with anonymous group labels

```
G ← GROUPBY(R, u(r))
for each report r_j in the fixed sender order do
    z_j ← ANONYMOUSLABEL(GROUP_OF(r_j, G))
    r′_j ← ATTACHGROUPTAG(r_j, z_j)
end for
prompt_true ← BUILDAGGREGATORPROMPT({r′_j}_{j=1}^J)
y_true ← AGGREGATOR(prompt_true)
return y_true
```

Figure D1: Algorithm D1: True Semantic Grouping.

Algorithm D2 Size-Matched Random Grouping

Require: Frozen reports R; true group-size multiset S_true; randomizations B = 20
Ensure: Predictions {y_rand^(b)} from all J reports in each randomization

```
for b = 1 to B do
    G_rand^(b) ← PARTITIONBYSIZES(SHUFFLE(R), S_true)
    for each report r_j in the original sender order do
        z_j^(b) ← ANONYMOUSLABEL(GROUP_OF(r_j, G_rand^(b)))
        r′_j^(b) ← ATTACHGROUPTAG(r_j, z_j^(b))
    end for
    prompt_rand^(b) ← BUILDAGGREGATORPROMPT({r′_j^(b)}_{j=1}^J)
    y_rand^(b) ← AGGREGATOR(prompt_rand^(b))
end for
return {y_rand^(b)}_{b=1}^B
```

Figure D2: Algorithm D2: Size-Matched Random Grouping.

Algorithm D3 Perfect Novelty Gate

Require: Receiver i; peer messages M_i^(t); receiver-specific seen set S_i^(t)
Ensure: Messages with repeated reasoning spans masked; Answer fields preserved

```
B ← [ ]
for each message m in M_i^(t) in the fixed sender order do
    u ← SEMANTICUNIT(REASONINGSPAN(m))
    if u is undefined then keep the reasoning span
    else if u □ S_i^(t) then add the reasoning span to B
    else keep the reasoning span; S_i^(t) ← S_i^(t) □ {u}
end for
M_i,nov^(t) ← MASKREASONING(M_i^(t), B)
Preserve all explicit Answer fields
return M_i,nov^(t)
```

Figure D3: Algorithm D3: Perfect Novelty Gate.

Algorithm D4 Matched Random Gate

Require: Novelty-blocked spans B_nov; candidate spans C
Ensure: Randomly blocked set B_rand matched in count and message features

```
Remove semantic-unit identities from the random selector
B_rand ← [ ]
for each blocked span b in B_nov do
    C_b ← MATCHEDCANDIDATES(b, C)
    c* ← MINCOSTMATCH(b, C_b)
    Append c* to B_rand; remove c* from C
end for
M_rand ← MASKREASONING(M, B_rand)
Preserve all explicit Answer fields
return M_rand
```

Figure D4: Algorithm D4: Matched Random Gate.

D.2 Input Transformations

Tables D1–D8 present paired examples of the model inputs under the four Experiment 3 interventions. True Semantic Grouping and Size-Matched Random Grouping operate at final aggregation and retain every agent report in its original order. The former assigns a shared anonymous group label to reports with the same controller-side semantic_unit_id; the latter draws a size-matched random partition without semantic provenance. Perfect Novelty Gate and Matched Random Gate operate during communication. Perfect Novelty Gate masks repeated reasoning spans according to receiver-specific semantic novelty, whereas Matched Random Gate masks a matched number of spans with matched message features but without semantic identity. Both preserve explicit Answer fields. The controller-side semantic_unit_id is never exposed directly to ordinary agents or the Central Aggregator.

Table D1: Aggregator input before True Semantic Grouping.

Question: Who is the artist of the "Freedom Suite"?

Agent 1
Answer: John Coltrane
Reasoning: John Coltrane is the artist of the "Freedom Suite".

Agent 2
Answer: John Coltrane
Reasoning: The "Freedom Suite" is attributed to John Coltrane as its artist.

Agent 3
Answer: John Coltrane
Reasoning: John Coltrane is the artist of the "Freedom Suite".

Agent 4
Answer: John Coltrane
Reasoning: John Coltrane is credited as the artist behind the "Freedom Suite".

Table D2: Aggregator input after True Semantic Grouping.

Question: Who is the artist of the "Freedom Suite"?

Agent 1
Answer: John Coltrane
Reasoning: John Coltrane is the artist of the "Freedom Suite". **[G1]**

Agent 2
Answer: John Coltrane
Reasoning: The "Freedom Suite" is attributed to John Coltrane as its artist. **[G1]**

Agent 3
Answer: John Coltrane
Reasoning: John Coltrane is the artist of the "Freedom Suite". **[G1]**

Agent 4
Answer: John Coltrane
Reasoning: John Coltrane is credited as the artist behind the "Freedom Suite". **[G1]**

Some reasoning statements contain anonymous group tags in square brackets.
Identical tags indicate that the corresponding reasoning statements share the same underlying semantic evidence unit.
The group tags do not indicate which answer is correct.

Table D3: Aggregator input before Size-Matched Random Grouping.

Question: Who is the artist of the "Freedom Suite"?

Agent 1
Answer: John Coltrane
Reasoning: John Coltrane is the artist of the "Freedom Suite".

Agent 2
Answer: John Coltrane
Reasoning: The "Freedom Suite" is attributed to John Coltrane as its artist.

Agent 3
Answer: John Coltrane
Reasoning: John Coltrane is the artist of the "Freedom Suite".

Agent 4
Answer: John Coltrane
Reasoning: John Coltrane is credited as the artist behind the "Freedom Suite".

Table D4: Aggregator input after Size-Matched Random Grouping.

Question: Who is the artist of the "Freedom Suite"?

Agent 1
Answer: John Coltrane
Reasoning: John Coltrane is the artist of the "Freedom Suite". **[G3C]**

Agent 2
Answer: John Coltrane
Reasoning: The "Freedom Suite" is attributed to John Coltrane as its artist. **[G3F]**

Agent 3
Answer: John Coltrane
Reasoning: John Coltrane is the artist of the "Freedom Suite". **[G3D]**

Agent 4
Answer: John Coltrane
Reasoning: John Coltrane is credited as the artist behind the "Freedom Suite". **[G3C]**

Some reasoning statements contain anonymous group tags in square brackets.
Identical tags indicate that the corresponding reasoning statements are assigned to the same proposed evidence group.
The group tags do not indicate which answer is correct.

Table D5: Peer-message input before Perfect Novelty Gate.

Question: Who is the artist of the "Freedom Suite"?

Receiver: Agent-01

Agent-02
Answer: David S. Ware
Reasoning: The artist of 'Freedom Suite' is David S. Ware, as stated in the assigned document.

Agent-03
Answer: John Coltrane
Reasoning: The assigned document states that John Coltrane is the artist of the 'Freedom Suite'.

Agent-04
Answer: UNKNOWN
Reasoning: The assigned document does not provide any information about the artist of the 'Freedom Suite' or any musical compositions. It only discusses the scientific name of the YEW tree.

Agent-05
Answer: David S. Ware
Reasoning: The artist of 'Freedom Suite' is David S. Ware, as stated in the assigned document.

Agent-06
Answer: John Coltrane
Reasoning: The assigned document states that John Coltrane is the artist of the 'Freedom Suite'.

Agent-07
Answer: UNKNOWN
Reasoning: The assigned document does not provide any information about the artist of the 'Freedom Suite'.

Agent-08
Answer: John Coltrane
Reasoning: The assigned document states that John Coltrane is the artist of the 'Freedom Suite'.

Agent-09
Answer: UNKNOWN
Reasoning: The assigned document does not provide any information about the artist of 'Freedom Suite' or any related information to answer the question.

Agent-10
Answer: John Coltrane
Reasoning: The assigned document states that John Coltrane is the artist of the 'Freedom Suite'.

Table D6: Peer-message input after Perfect Novelty Gate.

Question: Who is the artist of the "Freedom Suite"?

Receiver: Agent-01

Agent-02
Answer: David S. Ware
Reasoning: The artist of 'Freedom Suite' is David S. Ware, as stated in the assigned document.

Agent-03
Answer: John Coltrane
Reasoning: The assigned document states that John Coltrane is the artist of the 'Freedom Suite'.

Agent-04
Answer: UNKNOWN
Reasoning: The assigned document does not provide any information about the artist of the 'Freedom Suite' or any musical compositions. It only discusses the scientific name of the YEW tree.

Agent-05
Answer: David S. Ware
Reasoning: [omitted] [omitted] [omitted] [omitted] [omitted]

Agent-06
Answer: John Coltrane
Reasoning: [omitted] [omitted] [omitted] [omitted]

Agent-07
Answer: UNKNOWN
Reasoning: The assigned document does not provide any information about the artist of the 'Freedom Suite'.

Agent-08
Answer: John Coltrane
Reasoning: [omitted] [omitted] [omitted] [omitted]

Agent-09
Answer: UNKNOWN
Reasoning: The assigned document does not provide any information about the artist of 'Freedom Suite' or any related information to answer the question.

Agent-10
Answer: John Coltrane
Reasoning: [omitted] [omitted] [omitted] [omitted]

Table D7: Peer-message input before Matched Random Gate.

Question: Who is the artist of the "Freedom Suite"?

Receiver: Agent-01

Agent-02
Answer: David S. Ware
Reasoning: The artist of 'Freedom Suite' is David S. Ware, as stated in the assigned document.

Agent-03
Answer: John Coltrane
Reasoning: The assigned document states that John Coltrane is the artist of the 'Freedom Suite'.

Agent-04
Answer: UNKNOWN
Reasoning: The assigned document does not provide any information about the artist of the 'Freedom Suite' or any musical compositions. It only discusses the scientific name of the YEW tree.

Agent-05
Answer: David S. Ware
Reasoning: The artist of 'Freedom Suite' is David S. Ware, as stated in the assigned document.

Agent-06
Answer: John Coltrane
Reasoning: The assigned document states that John Coltrane is the artist of the 'Freedom Suite'.

Agent-07
Answer: UNKNOWN
Reasoning: The assigned document does not provide any information about the artist of the 'Freedom Suite'.

Agent-08
Answer: John Coltrane
Reasoning: The assigned document states that John Coltrane is the artist of the 'Freedom Suite'.

Agent-09
Answer: UNKNOWN
Reasoning: The assigned document does not provide any information about the artist of 'Freedom Suite' or any related information to answer the question.

Agent-10
Answer: John Coltrane
Reasoning: The assigned document states that John Coltrane is the artist of the 'Freedom Suite'.

Table D8: Peer-message input after Matched Random Gate.

Question: Who is the artist of the "Freedom Suite"?

Receiver: Agent-01

Agent-02
Answer: David S. Ware
Reasoning: [omitted] [omitted] [omitted] [omitted] [omitted]

Agent-03
Answer: John Coltrane
Reasoning: [omitted] [omitted] [omitted] [omitted]

Agent-04
Answer: UNKNOWN
Reasoning: The assigned document does not provide any information about the artist of the 'Freedom Suite' or any musical compositions. It only discusses the scientific name of the YEW tree.

Agent-05
Answer: David S. Ware
Reasoning: The artist of 'Freedom Suite' is David S. Ware, as stated in the assigned document.

Agent-06
Answer: John Coltrane
Reasoning: [omitted] [omitted] [omitted] [omitted]

Agent-07
Answer: UNKNOWN
Reasoning: The assigned document does not provide any information about the artist of the 'Freedom Suite'.

Agent-08
Answer: John Coltrane
Reasoning: [omitted] [omitted] [omitted] [omitted]

Agent-09
Answer: UNKNOWN
Reasoning: The assigned document does not provide any information about the artist of 'Freedom Suite' or any related information to answer the question.

Agent-10
Answer: John Coltrane
Reasoning: The assigned document states that John Coltrane is the artist of the 'Freedom Suite'.

E Additional Results and Statistical Analysis

E.1 Cross-Model Replication

Within the same controlled E1 setting, the surface-repetition pattern also appears in two additional models. As M_w increases from 1 to 8, GPT-4o-mini accuracy decreases from 95% to 70% while target-wrong selection increases from 2% to 29%. For Mistral-7B-Instruct-v0.3, accuracy decreases from 82% to 65% and target-wrong selection increases from 5% to 30%, following the integer percentages displayed in Figure E1. These results provide a bounded cross-model replication under the tested models and controlled dataset; they do not establish generality to other architectures, retrieval settings, or open-web evidence.

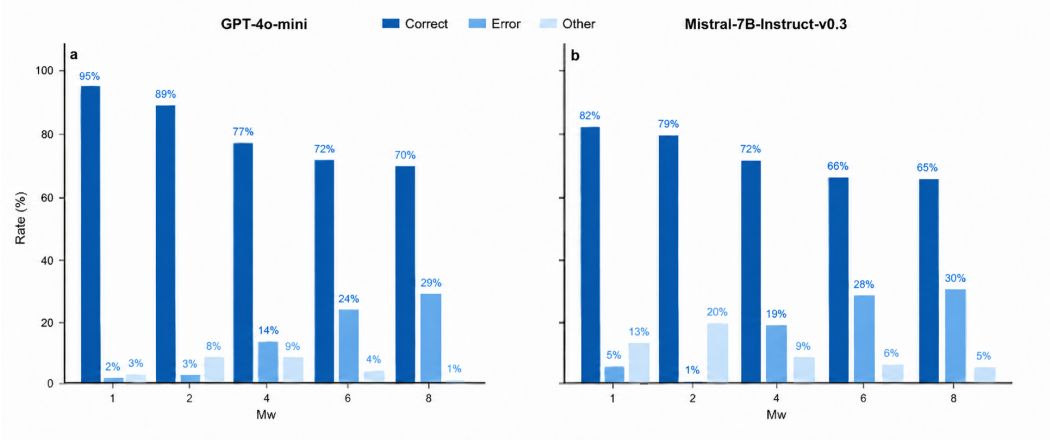


Figure E1: Cross-model E1 results under increasing surface repetition.

E.2 Incremental Effect of Genuine Peer Content

Figure E2 decomposes the incremental contribution of genuine peer content to the final target-wrong advantage after two communication rounds. With the continuity-corrected group-level hard-answer log-odds b = log[(W + 0.5)/(C + 0.5)], Equation (E.1) gives the compact definitions of $A_{peer}(M_w)$ and $DID_{peer}(m_a, m_b)$ with the same subtraction order as the main text; Equation (E.2) expands the DID contrast into its Full Debate and Matched Context components. The upper row shows A_{peer} at adjacent repetition levels and their pairwise contrasts; the lower row expands each contrast into its four signed log-odds components. The adjacent DID values are +1.894 (1→2), -1.163 (2→4), -0.275 (4→6), and -0.293 (6→8). They reveal a non-monotonic, state-dependent communication effect rather than a steadily increasing amplification process.

$$A_{\text{peer}}(M_w) = b_{\text{Full}}(M_w) - b_{\text{Matched}}(M_w), \quad DID_{\text{peer}}(m_a, m_b) = A_{\text{peer}}(m_a) - A_{\text{peer}}(m_b) \tag{E.1}$$

$$DID_{\text{peer}}(m_a, m_b) = [b_{\text{Full}}(m_a) - b_{\text{Matched}}(m_a)] - [b_{\text{Full}}(m_b) - b_{\text{Matched}}(m_b)] \tag{E.2}$$

A_{peer} is positive for $M_w \in \{2, 4, 6, 8\}$ but is non-monotonic. The main-text high-versus-low contrast is $DID_{\text{peer}}(8, 1) = 0.163$, with a 95% paired-bootstrap interval of $[-0.980, 1.337]$. Because the interval includes zero, this contrast is a descriptive diagnostic rather than evidence of monotonic amplification. The bootstrap resamples base_qa_id clusters, retains paired Full Debate and Matched Context outcomes across all M_w levels, and uses 2,000 replicates. The estimand is based on group-level hard-answer counts, not token probabilities or an error-retention rate.

Table E1: Corrected final W/C log-odds, A_{peer}, and question-level paired-bootstrap intervals.

M_w	$b_{\text{Matched}}(M_w)$	$b_{\text{Full}}(M_w)$	A_{peer}	**95% paired bootstrap interval**
1	-3.160	-3.219	-0.059	$[-1.178, 1.054]$
2	-3.085	-1.249	$+1.835$	$[1.083, 3.609]$
4	-1.441	-0.769	$+0.672$	$[0.261, 1.179]$
6	-0.648	-0.251	$+0.397$	$[0.062, 0.760]$
8	-0.441	-0.336	$+0.105$	$[-0.235, 0.447]$

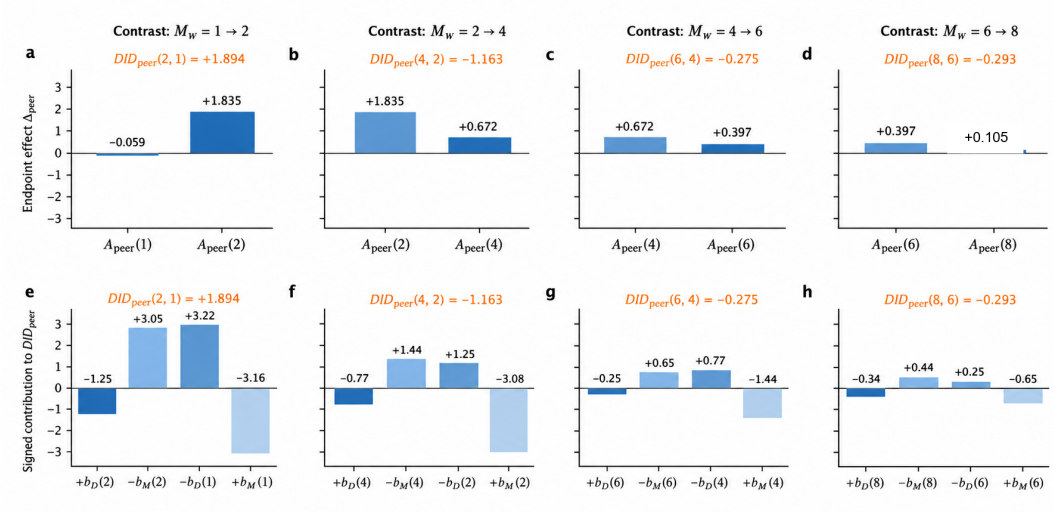


Figure E2: Decomposition of DID_{peer} across adjacent repetition-level contrasts.