SUPPORT-FLOW CONTROL RESISTS PERSUASION BY ECHO IN FIXED-EVIDENCE MULTI-AGENT DEBATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent debate can corrupt its own evidential record: once a weak claim is reused as a premise, later turns amplify its authority without adding evidence. We formalize this failure as *persuasion by echo* and introduce *support-flow control*, which governs how evidential credit propagates between claims while an argument is being written. RelayGuard instantiates this principle by decomposing direct, inherited, and unsupported support, weighting unsupported reuse by downstream influence, and expiring influential claims until new evidence re-grounds them. Under matched evidence and generation budgets, RelayGuard attains a 26.7% judge-fooled rate across three controlled tasks, versus 29.0% for active ATE-FLAT, 29.3% for Answer+Verifier, and 29.8% for SLE-FLAT. The advantage extends to two fixed-corpus evaluations: RelayGuard lowers judge-fooled rate by 2.5 points on FEVER-EVIDENCE and 2.8 points on ContractNLI relative to Answer+Verifier. Together, matched controls, mechanism diagnostics, and external fixed-corpus results establish support inheritance as a controllable state of multi-agent reasoning and transcript formation as a decisive point for provenance-aware oversight.

1 INTRODUCTION

Debate is often evaluated through the evidence attached to each turn, yet a multi-turn argument has a second provenance structure: the path by which one claim lends authority to another. A weak assertion can be reused as a premise, inherited by downstream claims, and amplified until repetition substitutes for grounding. We call this failure *persuasion by echo*. We study this failure within debate-based supervision (Irving et al., 2018; Michael et al., 2023; Khan et al., 2024). This failure matters wherever agents construct an auditable argument over a common record, including fact checking, evidence-grounded question answering, document review, and scalable oversight.

This paper introduces *support-flow integrity* as an operational objective for such systems. Source attribution records where a statement points; support-flow integrity measures whether authority passed through the transcript remains anchored to that record. The distinction reveals a consequential intervention point. Local citation improves source attachment, completed-answer verification improves selection, and judge-facing provenance improves the final audit. Support-flow control acts during generation, when a debater's next move can still change both the dependency path and the evidence carried forward.

RelayGuard is a closed-loop controller for that objective. Each structured claim declares evidence and parent claims. A claim-conditioned decomposition then separates direct, inherited, and unsupported support, while dependency centrality measures how much downstream argument a claim can influence. RelayGuard exposes this state to both debaters, prices repeated unsupported inheritance by influence, and expires stale high-impact claims until new evidence re-grounds them. The resulting quantity—unsupported influence mass—connects local evidence coverage to transcript-level authority and turns provenance into a control signal for subsequent generation.

The experiments test the control principle against its closest alternative explanations. On three controlled tasks, RelayGuard reduces macro judge-fooled rate from 41.4% to 26.7% and leads an activity-matched non-graph controller by 2.3 points, a token-matched Answer+Verifier pipeline by 2.6, and SLE-FLAT by 3.1. RelayGuard also doubles post-expiration re-grounding relative to active ATE-FLAT, connecting graph-aware enforcement to the availability of new evidence. The decision

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.

advantage persists on fixed FEVER-EVIDENCE and ContractNLI evaluations with frozen controller settings.

The paper makes three contributions:

- We formulate support-flow integrity for fixed-evidence debate and operationalize persuasion by echo through claim-conditioned inherited support and unsupported influence mass.
- We develop RelayGuard, a debater-facing controller that couples dependency-aware support state, influence-weighted reuse cost, and evidence-triggered expiration in a deterministic turn-level loop.
- We establish the mechanism and decision consequence with matched active and verifier controls, a fully specified factorial comparison, item-level uncertainty, and fixed-corpus evaluation on FEVER-EVIDENCE and ContractNLI.

2 Related Work: From Source Attribution to Support-Flow Control

Grounding and revision. Citation-aware generation and retrieval-based revision attach evidence to outputs (Nakano et al., 2021; Gao et al., 2023); question decomposition improves reasoning faithfulness (Radhakrishnan et al., 2023). RelayGuard uses the same local grounding interface as an input to a different control object: the evidential credit that propagates from earlier claims into later ones. Claim-conditioned normalization evaluates each claim's direct support against its required evidence before that support can pass through parent edges.

Critique and verification. Self-critique, factuality scoring, process supervision, and trained verifiers evaluate or refine answers (Saunders et al., 2022; Madaan et al., 2023; Min et al., 2023; Lightman et al., 2023; Cobbe et al., 2021). Model-written evaluations expose additional behavioral failures (Perez et al., 2022). These methods motivate the strong Answer+Verifier and reranked-verifier comparators in our evaluation. RelayGuard complements their output-selection objective with a generation-time objective: keep the support chain grounded as claims are composed.

Debate protocols and agent graphs. Multi-agent debate and prover–verifier games shape interaction, information exchange, and final legibility (Du et al., 2023; Liang et al., 2023; Chan et al., 2023; Kirchner et al., 2024; Parrish et al., 2022; Brown-Cohen et al., 2023). Graph-structured agent systems optimize communication topology (Zhuge et al., 2024). RelayGuard instead uses the claim-dependency graph as a live provenance state: edges identify inherited evidential credit, influence identifies consequential unsupported reuse, and enforcement requires new evidence before that credit can propagate again. This coupling of support decomposition, dependency influence, and debater-facing re-grounding defines the support-flow control problem tested below.

3 Support-Flow Integrity in Fixed-Evidence Debate

Debate (Irving et al., 2018) is often motivated as a way for a weaker judge to supervise a stronger system (Amodei et al., 2016; Christiano et al., 2018; Leike et al., 2018; Bowman et al., 2022). Under ideal conditions, the truthful side can expose falsehoods through adversarial exchange (Brown-Cohen et al., 2023). In practice, LLMs serve as debaters and LLMs or humans as judges (Michael et al., 2023; Khan et al., 2024). We study evidence-grounded factual disputes whose correctness is determined by a finite record shared by both debaters, making the support passed between claims observable and adjudicable.

Our debate setting pairs question q and evidence corpus $\mathcal{E} = \{e_1, \ldots, e_n\}$ with two agents, $A_{\text{pro}}, A_{\text{con}}$, and a judge, J. The protocol requires the agent acting on turn t to output the following *operational structured claim record* $c_t = \big(\sigma_t,\ \tau_t,\ S(c_t),\ P(c_t)\big)$: σ_t supplies the surface utterance as a natural-language string, $\tau_t \in \{\text{ASSERTION}, \text{REBUTTAL}, \text{CITATION}, \text{SYNTHESIS}\}$ specifies its type, $S(c_t) \subseteq \mathcal{E}$ gives its supporting evidence, and $P(c_t) \subseteq \{c_1, \ldots, c_{t-1}\}$ lists its parents, possibly none. This tuple occupies a required output slot in the protocol: decoding constrains agents to produce a schema-conforming fenced JSON block, while the deterministic parser in Appendix I rejects a non-conforming turn and permits one retry; the locked parser audit covers 61,200 multi-turn

debates on the primary tasks (734,400 agent turns), with 5.6% of turns succeeding on that retry and persistent failures (0.9%) recorded with defaults $S = \emptyset, P = \emptyset$. This turn-level interface exposes support state for intervention and audit. Retrieval identifies the relevant subset $\mathcal{E}_{\text{rel}} \subseteq \mathcal{E}$; within it, $\mathcal{E}_{\text{req}}(c) \subseteq \mathcal{E}_{\text{rel}}$ operationally approximates the *claim-required* subset for c by selecting snippets that the fixed backend, conditioned on σ_t, treats as necessary to license the claim (BGE-large-en-v1.5 top-k with $k = 4$ and a fixed 0.55 cosine floor; $|\mathcal{E}_{\text{req}}(c)| \leq 4$).

For a claim c to count as repeating an earlier claim c', c must convey similar meaning to c' *and* add no evidence beyond that of c':

$$\text{rep}(c, t) = \big|\{c' \in \{c_1, \ldots, c_{t-1}\} : \text{sim}(c, c') > \delta \ \wedge \ S(c) \subseteq S(c')\}\big|. \tag{1}$$

With the subset relation directed as $S(c) \subseteq S(c')$, adding strictly new evidence excludes a claim from the repetition count, as intended by the prose definition. We compute this quantity exactly once, when c is emitted, and cache $r(c) = \text{rep}(c, t_0(c))$. Thus the comparison set for $r(c)$ contains only claims preceding c and never c itself. Our pilot covered 500 standard-debate transcripts: when judges chose the factually incorrect answer, 31.4% of all claims were unsupported echoes (satisfying $\text{rep}(c, t) > 0$ and $|S(c)| = 0$), compared with 12.8% when decisions were correct; this difference of 18.6 pp motivates targeted intervention.

Persuasion by echo is the semantic restatement of weakly grounded claims that becomes influential as support accumulates through downstream reuse. The control objective is to keep that inherited authority proportional to evidence while preserving useful restatement of grounded claims. The evaluated regime has a finite evidence bundle, evidence-resolvable correctness, and adversarially reusable claims; Section 5 tests controlled and externally fixed instances of this regime.

4 RelayGuard: Closed-Loop Support-Flow Control

RelayGuard implements a three-stage feedback loop: estimate the support carried by each claim, locate unsupported mass with high downstream influence, and intervene before the next claims are generated. The ledger provides claim-level state, dependency edges preserve support transfer, centrality prioritizes consequential paths, and tax plus expiration translate that state into immediate incentives. Re-grounding closes the loop by restoring a claim only when its evidence set expands.

We represent the transcript's support-flow state with unsupported influence mass. We decompose support into three components for a claim c on turn t. The clipping notation is $[x]_0^1 = \min(\max(x, 0), 1)$:

$$d(c) = \left[\frac{|S(c) \cap \mathcal{E}_{\text{req}}(c)|}{\max(1, |\mathcal{E}_{\text{req}}(c)|)}\right]_0^1, \tag{2}$$

$$\tilde{t}(c) = \begin{cases} 0 & \text{if } P(c) = \emptyset, \\ \frac{1}{|P(c)|} \sum_{c' \in P(c)} a_t(c') d(c') \cdot \mathbb{1}[r(c') = 0] & \text{otherwise,} \end{cases} \tag{3}$$

$$t(c) = \min\big(\tilde{t}(c),\ 1 - d(c)\big), \qquad u(c) = 1 - d(c) - t(c). \tag{4}$$

Instead of normalizing uniformly by $1/|\mathcal{E}_{\text{rel}}|$, we make $d(c)$ *claim-conditioned*: citing both snippets needed by a two-snippet claim yields $d(c) = 1$, with no penalty for unrelated snippets. The deterministic retrieval pipeline computes $\mathcal{E}_{\text{req}}(c)$ from σ_t and is shared by all compared protocols, preventing selective omission from changing the support criterion. Every claim has a valid decomposition, including those without parents ($P(c) = \emptyset \Rightarrow t(c) = 0, u(c) = 1 - d(c)$); its components are always non-negative ($d, t, u \geq 0$) and sum to one ($d + t + u = 1$). Here $a_t(c') \in \{0, 1\}$ is the claim's post-expiration active-state indicator at turn t. Together, the two indicators block transitive support from parents that have expired or were tagged as echoes when they themselves were emitted. We write $\text{cent}(c)$ for claim c's PageRank centrality, computed on $\mathcal{G} = (\mathcal{C}, \mathcal{P})$, the claim graph, using damping $d_{\text{pr}} = 0.85$ (Page et al., 1998). The *unsupported influence mass* of the graph at turn t is

$$U(t) = \sum_{c \in \mathcal{C}_t} \text{cent}(c) \cdot u(c). \tag{5}$$

Higher $U(t)$ means that more of the transcript's current influence is carried by poorly grounded claims. RelayGuard transforms this state into three enforcement rules that target influential claims inheriting support without fresh evidence.

Our first mechanism, the shared evidence ledger $\mathcal{L}$, stores each claim's (d, t, u) values after each turn. Judge-only presentation makes this state available at decision time; RelayGuard also exposes it to the debaters while they can still alter the support chain. The resulting common state makes unsupported inheritance strategically legible to both sides and supplies the input to active enforcement.

The second mechanism is a centrality-weighted endorsement tax:

$$\mathrm{tax}(c,t) = \alpha \cdot \mathrm{cent}(c) \cdot \frac{\mathrm{rep}(c,t)}{1+d(c)}, \qquad T_{\mathrm{cum}}(A_i) = \sum_{c \in \mathcal{C}_i} \mathrm{tax}(c,T). \tag{6}$$

Dividing by $1 + d(c)$ lowers the cost of repeating well-evidenced claims, reflecting the smaller epistemic harm of restating a directly supported conclusion compared with echoing an assertion lacking a source. The judge receives both the transcript and the accumulated tax.

The third mechanism is claim expiration. For claim c, whose introduction occurs on turn t_0, *expiration* takes effect at turn t's end if $t - t_0 \geq N_{\mathrm{exp}}$ and $\mathrm{cent}(c) > \theta$. Expiration sets $a_t(c) = 0$: the claim remains in the immutable ledger but is excluded from the active graph and therefore has zero centrality and contributes zero inherited support. A later claim c_t re-grounds an inactive parent $c \in P(c_t)$ precisely when $S(c_t) \setminus S(c) \neq \emptyset$; this restores $a_t(c) = 1$ before the active-graph PageRank update. In our experiments we set $\theta = 0.3$ and $N_{\mathrm{exp}} = 3$.

Turn t follows four phases in a fixed order: (P1) generate c_t and parse it; (P2) add c_t, restore parents that meet the re-grounding predicate, and converge PageRank on the graph induced by active claims (ℓ_1 tolerance 10^{-6}, typically ≤ 30 iterations); (P3) evaluate the inclusive expiration predicate, zeroing the affected active indicators and centralities; (P4) run PageRank again on the active graph that remains, calculate $d, \tilde{t}, t, u, r, \mathrm{tax}$, and $U(t)$, then store these in $\mathcal{L}$. This ordering makes the active graph, centralities, and inherited-support calculation a single deterministic state transition given the transcript.

Semantic deduplication uses open-weight BGE-large-en-v1.5 embeddings served through `sentence-transformers` 2.7.0 (1024-d, FP16). We measure similarity by $\mathrm{sim}(c, c') = \cos\big(\mathrm{BGE}(\sigma_c), \mathrm{BGE}(\sigma_{c'})\big)$; the threshold $\delta = 0.85$ maximizes Youden's J on 400 held-out claim pairs, which two authors annotated independently (Cohen's $\kappa = 0.81$; ROC AUC $= 0.93$). This embedding model also determines $\mathcal{E}_{\mathrm{req}}(c)$; stability under nearby choices of (α, θ, δ) is shown in Appendix H. RelayGuard is deterministic given the transcript and seed (Appendix A).

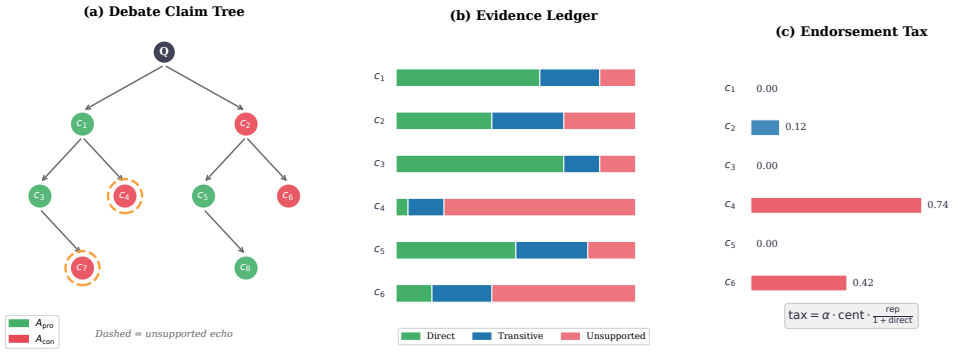

Figure 1: RelayGuard closes the support-flow loop before adjudication. *Left*: dependency edges expose inherited support paths. *Center*: claim-conditioned state separates direct, inherited, and unsupported mass. *Right*: influence-weighted reuse cost and expiration turn high-impact unsupported reuse into a re-grounding action. Enlarged panels appear in Appendix Q.

Figure 1 shows the state–action coupling. Dependency edges identify which earlier claims license a new statement, the ledger tests whether the inherited credit reaches evidence, and influence-weighted enforcement changes the next available action before repetition dominates the record. PageRank supplies an efficient influence estimate; the closed loop converts claim-conditioned provenance into debater behavior.

5 EVALUATION

5.1 PROTOCOL, COMPARATORS, AND UNCERTAINTY

The controlled suite contains Multi-hop Evidence QA (500 questions) following Yang et al. (2018), False-Lemma Persuasion (200 controlled scenarios), and Legal Snippet Comparison (150 case pairs). External evaluation fixes the public corpus, evidence, and labels for FEVER-EVIDENCE (Thorne et al., 2018) and ContractNLI (Koreeda & Manning, 2021). Hyperparameters are selected on a disjoint 250-item calibration split and then frozen across evaluation sets.

Debaters are GPT-4o (`2024-08-06`), Llama-3.1-70B-Instruct (Grattafiori et al., 2024), and Qwen-2.5-72B-Instruct (Qwen et al., 2024). Open-weight models run with vLLM 0.5.4 (Kwon et al., 2023), FP16, tensor parallelism 4, on four NVIDIA H100 SXM5 80GB GPUs (CUDA 12.4; driver 550.54.15). Each debate has six pro/con rounds (12 agent turns), shared evidence, debater temperature 0.7, and judge temperature 0.0. The primary judge is GPT-4o (`2024-08-06`); transfer checks use Claude-3.5-Sonnet (`2024-10-22`), Gemini-2.0-Flash (`2025-02-05`), and an open Qwen judge.

Comparators cover Standard Debate, Citation-Only, Citation+Repetition Penalty, Retrieve-and-Critique, Judge-only Ledger, Ledger-only, matched-information SLE-FLAT, active temporal ATE-FLAT, and a matched-budget Answer+Verifier best-of-4 pipeline. All share evaluation items, evidence access, and seed set $\{13, 29, 41, 53, 67, 79, 89, 97\}$; the closest primary comparators match realized token budgets within $\pm 5\%$, while Standard and broader stress tests have the costs reported in the appendix. Across the primary debate-style blocks, we analyze $88{,}400$ runs. The $61{,}200$-debate parser audit is the structured multi-turn subset of this comparison total.

Judge-fooled rate is the fraction of adjudications selecting the factually wrong side under the gold key. It directly measures the target decision: whether changing transcript construction improves final evidence-grounded adjudication while the record is held fixed. In the primary and public-benchmark outcome tables, each x$\pm$y gives the point estimate and 95% percentile-bootstrap interval half-width from 1,000 item-stratified resamples that retain all eight paired seed outcomes. Method contrasts resample the same item–seed clusters jointly; the inferential unit is the underlying item. Tables explicitly labelled as resampling per-seed means use the seed mean as the resampling unit; their intervals are not item-level intervals.

Table 1: Primary-suite headline comparison. Judge-fooled rate (%, ↓) under the GPT-4o judge; eight seeds and 95% item-stratified bootstrap intervals. ATE-FLAT uses active temporal matching and is evaluated at macro level.

Method	MH-QA	False-Lemma	Legal	Macro
Std. debate	38.4±1.3	44.2±1.7	41.6±1.5	41.4
Cit.-only	32.8±1.2	38.6±1.5	35.2±1.4	35.5
Cit.+rep.	30.1±1.1	33.4±1.3	32.3±1.2	31.9
Ret.+crit.	28.5±1.0	32.1±1.3	30.8±1.2	30.5
Ans.+ver.	27.8±1.0	31.2±1.2	29.0±1.1	29.3
SLE-FLAT	27.8±1.0	31.6±1.2	30.0±1.1	29.8
ATE-FLAT	–	–	–	29.0±1.1
RelayGuard	24.6±0.9	28.7±1.1	26.9±1.0	26.7

5.2 PRIMARY OUTCOME: SUPPORT-FLOW CONTROL BEATS MATCHED ALTERNATIVES

Table 1 shows a 14.7-point reduction from Standard Debate and persistent advantages over the strongest nearby alternatives: 2.3 points versus active ATE-FLAT, 2.6 versus Answer+Verifier, and 3.1 versus SLE-FLAT. RelayGuard is best on every controlled task. Appendix C retains the explicitly specified 2^3 controller configurations; we do not treat incompatible duplicate deletion rows as a separate ablation result.

5.3 Mechanism Isolation: Dependency-Aware Enforcement Changes the Transcript

Judge-only provenance reaches 37.3% macro fooled rate and debater-visible ledger-only reaches 33.2%, demonstrating gains from both intervention timing and enforcement. Table 2 separates ledger visibility, local repetition control, dependency-aware tax, and expiration. The active ATE-FLAT control has similar mean expiration activity (2.7 versus 3.1 events per debate) yet reaches 30.6% post-expiration re-grounding, compared with 63.1% for RelayGuard, and trails by 2.3 points. Together, these subtractions show that dependency-aware targeting converts similar temporal activity into more new evidence and fewer final errors.

Table 2: Mechanism isolation on the primary suite (macro judge-fooled rate, ↓). J-L and D-L denote judge- and debater-visible ledgers; Local, Graph, and Exp. denote local repetition control, dependency-aware tax, and claim expiration. SLE-FLAT is the passive matched-information comparator and has no expiration predicate.

Variant	J-L	D-L	Local	Graph	Exp.	Macro	Δ
Std. debate	–	–	–	–	–	41.4±1.4	0.0
Cit.-only	–	–	–	–	–	35.5±1.3	−5.9
Cit.+rep.	–	–	Yes	–	–	31.9±1.2	−9.5
Ans.+ver.	n/a	n/a	n/a	n/a	n/a	29.3±1.1	−12.1
Judge-only ledger	Yes	–	–	–	–	37.3±1.3	−4.1
Ledger-only	Yes	Yes	–	–	–	33.2±1.2	−8.2
Ledger+tax	Yes	Yes	–	Yes	–	29.8±1.1	−11.6
Ledger+exp.	Yes	Yes	–	–	Yes	31.5±1.2	−9.9
SLE-FLAT	Yes	Yes	Yes	–	–	29.8±1.1	−11.6
RelayGuard	Yes	Yes	–	Yes	Yes	26.7±1.0	−14.7

5.4 External Fixed-Corpus Evaluation

FEVER-EVIDENCE uses 3,000 claim-level adjudication cells with gold Wikipedia evidence. ContractNLI contributes a 607-contract composite over its 17 fixed hypotheses per contract (Appendix O). Corpus, evidence, labels, and RelayGuard hyperparameters remain fixed; both families preserve the closest-comparator ordering.

Table 3: Fixed-corpus evaluation on FEVER-EVIDENCE. Judge-fooled rate (%, ↓) under the GPT-4o judge; eight seeds and 95% item-stratified bootstrap intervals. Overall point estimates are item-count-weighted means of the displayed length strata ($n = 1{,}074/1{,}248/678$), rounded after aggregation.

FEVER-EVIDENCE (n=3,000)				
Method	Overall	Short	Med.	Long
Std. debate	33.3	29.6±1.3	34.2±1.2	37.3±1.5
Cit.-only	28.6	25.7±1.1	29.4±1.1	32.1±1.3
Cit.+rep.	26.6	23.1±1.0	26.8±1.0	29.5±1.2
Ans.+ver.	24.2	21.6±1.0	24.7±1.0	27.6±1.2
Ledger-only	25.4	22.3±1.0	25.9±1.0	28.1±1.2
SLE-FLAT	24.5	21.4±1.0	24.9±1.0	27.4±1.2
RelayGuard	21.7	19.4±0.9	22.1±0.9	24.5±1.1

RelayGuard is best on every FEVER slice. It lowers the stratum-weighted FEVER rate relative to Answer+Verifier by 2.5 points and relative to SLE-FLAT by 2.8 points. ContractNLI shows corresponding 2.8- and 2.2-point advantages at the contract-composite level (Appendix O). Longer evidence chains preserve the FEVER advantage, directly testing the regime in which inherited dependencies are most consequential.

Table 4: Closest-comparator matrix. Entries are judge-fooled rates (%, ↓); ± values are 95% interval half-widths. Tokens and minutes are measured on the primary calibration split. Answer+Verifier+Reranker uses best-of-4 generation, independent evidence-aware verifier scoring, and a separate aggregator without RelayGuard features.

Method	Primary	FEVER	Contract	Tokens	Min.
RelayGuard	26.7±1.0	21.7	21.5	6,172	4.60
ATE-FLAT	29.0±1.1	23.9 ± 0.8	23.2 ± 0.9	6,128	4.48
Ans.+Verifier	29.3	24.2	24.3	6,089	4.26
Ans.+Ver.+Rerank	28.5 ± 1.0	23.6 ± 0.8	22.9 ± 0.9	6,341	4.74

Table 4 strengthens both subtraction tests: RelayGuard leads the active non-graph controller on every external family and also leads a stronger reranked verifier while remaining within the same realized token envelope. The external point estimates use the aggregation rules stated with their corresponding benchmark tables.

5.5 Stress Tests

Under combined adaptive attack, evidence collusion, and judge-only-transcript evaluation, RelayGuard remains the best tested method at 31.8%, 30.6%, and 34.8% fooled rate, compared with 36.4%, 33.1%, and 37.1% for Answer+Verifier (Appendix B). These stress tests preserve the design advantage under substantially degraded evidence and attack conditions.

6 Conclusion

Multi-agent reasoning carries evidence through a transcript as well as attaching it to individual statements. RelayGuard makes that support flow measurable and governable: claim-conditioned decomposition exposes unsupported inheritance, dependency influence identifies consequential reuse, and expiration converts stale authority into a demand for new evidence. The controller lowers primary-suite error by 14.7 points from Standard and by 2.3–3.1 points from active ATE-FLAT, Answer+Verifier, and SLE-FLAT; its $2.1\times$ re-grounding increase connects the outcome advantage to a higher rate of evidence-restoring action. FEVER-EVIDENCE and ContractNLI preserve the closest-comparator ordering under fixed corpora and frozen controller settings. Support-flow integrity is therefore a practical design objective for provenance-aware oversight: govern how generators propagate evidential credit while the argument is formed.

AI Use Statement

Generative AI was used only for manuscript writing, editing, and presentation. The authors reviewed all edits and remain responsible for the final content.

Reproducibility Statement

Sections 4–5 specify the control rule, models, inference settings, comparators, splits, seeds, and uncertainty procedures. The appendix gives the executable transition order, calibration grid, budget matching, external-evaluation scope, stress-test protocols, and active-control diagnostics used to specify the reported analyses.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*, 2016. doi: 10.48550/arXiv.1606.06565. URL https://arxiv.org/abs/1606.06565.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring Progress on Scalable

Oversight for Large Language Models. *arXiv preprint arXiv:2211.03540*, 2022. doi: 10.48550/arXiv.2211.03540. URL https://arxiv.org/abs/2211.03540.

Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. Scalable AI Safety via Doubly-Efficient Debate. *arXiv preprint arXiv:2311.14125*, 2023. doi: 10.48550/arXiv.2311.14125. URL https://arxiv.org/abs/2311.14125.

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. doi: 10.48550/arXiv.2308.07201. URL https://arxiv.org/abs/2308.07201.

Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018. doi: 10.48550/arXiv.1810.08575. URL https://arxiv.org/abs/1810.08575.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

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. doi: 10.48550/arXiv.2305.14325. 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, 2023. doi: 10.18653/v1/2023.acl-long.910. URL https://aclanthology.org/2023.acl-long.910/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018. doi: 10.48550/arXiv.1805.00899. URL https://arxiv.org/abs/1805.00899.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with More Persuasive LLMs Leads to More Truthful Answers. *arXiv preprint arXiv:2402.06782*, 2024. doi: 10.48550/arXiv.2402.06782. URL https://arxiv.org/abs/2402.06782.

Jan Hendrik Kirchner, Yining Chen, Harri Edwards, Jan Leike, Nat McAleese, and Yuri Burda. Prover-Verifier Games improve legibility of LLM outputs. *arXiv preprint arXiv:2407.13692*, 2024. doi: 10.48550/arXiv.2407.13692. URL https://arxiv.org/abs/2407.13692.

Yuta Koreeda and Christopher D. Manning. ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1907–1919, 2021. doi: 10.18653/v1/2021.findings-emnlp.164. URL https://aclanthology.org/2021.findings-emnlp.164/.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018. doi: 10.48550/arXiv.1811.07871. URL https://arxiv.org/abs/1811.07871.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. *arXiv preprint arXiv:2305.19118*, 2023. doi: 10.48550/arXiv.2305.19118. URL https://arxiv.org/abs/2305.19118.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050*, 2023. doi: 10.48550/arXiv.2305.20050. URL https://arxiv.org/abs/2305.20050.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R. Bowman. Debate Helps Supervise Unreliable Experts. *arXiv preprint arXiv:2311.08702*, 2023. doi: 10.48550/arXiv.2311.08702. URL https://arxiv.org/abs/2311.08702.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. doi: 10.48550/arXiv.2112.09332. URL https://arxiv.org/abs/2112.09332.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Libraries, 1998. URL https://theory.stanford.edu/~rajeev/papers.html.

Alicia Parrish, Harsh Trivedi, Ethan Perez, Angelica Chen, Nikita Nangia, Jason Phang, and Samuel R. Bowman. Single-Turn Debate Does Not Help Humans Answer Hard Reading-Comprehension Questions. *arXiv preprint arXiv:2204.05212*, 2022. doi: 10.48550/arXiv.2204.05212. URL https://arxiv.org/abs/2204.05212.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv preprint arXiv:2212.09251*, 2022. doi: 10.48550/arXiv.2212.09251. URL https://arxiv.org/abs/2212.09251.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Question Decomposition Improves the Faithfulness of Model-Generated Reasoning. *arXiv preprint arXiv:2307.11768*, 2023. doi: 10.48550/arXiv.2307.11768. URL https://arxiv.org/abs/2307.11768.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022. doi: 10.48550/arXiv.2206.05802. URL https://arxiv.org/abs/2206.05802.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

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/.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. GPTSwarm: Language Agents as Optimizable Graphs. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62743–62767. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhuge24a.html.

A Protocol Algorithm

Algorithm 1 RelayGuard support-flow control (one full debate)

Require: Question q, evidence corpus $\mathcal{E}$, agents $A_{\text{pro}}, A_{\text{con}}$, judge J, total agent turns $T = 2R$ (with R pro/con rounds), tax coefficient α, centrality threshold θ, expiration window N_{exp}, similarity threshold δ, embedding model BGE

1: Initialize ledger $\mathcal{L} \leftarrow \emptyset$, claim graph $\mathcal{G} \leftarrow (\emptyset, \emptyset)$
2: **for** $t = 1$ **to** T **do**
3: $A_i \leftarrow$ acting agent at turn t
4: (P1 Emit & Parse) Sample agent output $o_t \leftarrow A_i.\textsc{Act}(q, \mathcal{L}, \mathcal{G}, \mathcal{E})$; parse fenced JSON block via $\textsc{Parse}(o_t) \rightarrow (\sigma_t, \tau_t, S(c_t), P(c_t))$; on parse failure, retry once with a structured-output reminder; log persistent failures and default to $S = \emptyset, P = \emptyset$
5: (P2 Insert, Re-ground & Centrality) Insert c_t and parent edges. For each inactive $c \in P(c_t)$ with $S(c_t) \setminus S(c) \neq \emptyset$, set $a_t(c) \leftarrow 1$; set $a_t(c_t) \leftarrow 1$. Run PageRank on the graph induced by active claims to convergence (ℓ_1 tol. 10^{-6}, $d_{\text{pr}} = 0.85$), updating $\text{cent}(c)$ for active c.
6: (P3 Expire) For every active c with $\text{cent}(c) > \theta \wedge t - t_0(c) \geq N_{\text{exp}}$, set $a_t(c) \leftarrow 0$ and $\text{cent}(c) \leftarrow 0$.
7: (P4 Score & Tax) Recompute PageRank on the resulting active graph; compute $\mathcal{E}_{\text{req}}(c_t)$ via BGE-retrieval on σ_t, cache $r(c_t) = \text{rep}(c_t, t)$ (Eq. 1), then compute $d, \tilde{t}, t, u$ (Eqs. 2–4), $\text{tax}(c_t, t)$ (Eq. 6), and $U(t)$ (Eq. 5); commit to $\mathcal{L}$.
8: **end for**
9: Compute $T_{\text{cum}}(A_{\text{pro}}), T_{\text{cum}}(A_{\text{con}})$
10: $\hat{a} \leftarrow J.\textsc{Judge}(q, \text{transcript}, \mathcal{L}, T_{\text{cum}})$
11: **return** $\hat{a}$

Algorithm 1 fixes the state transition that makes support-flow enforcement deterministic. Parsing and scoring occur every turn, expiration follows a pre-expiration active-graph update and precedes the post-expiration update and ledger commitment, and persistent parse failure receives the conservative empty-support state. Consequently, an inactive or malformed parent cannot propagate evidential credit during the same turn.

B Adversarial and Adaptive Attacks

Static attacks. *Thin-snippet gaming*: attaches minimally relevant snippets that share surface keywords but provide no substantive support. *Repetition tactics*: paraphrases the core claim extensively to stay below δ while maintaining semantic equivalence.

Table 5: Judge-fooled rate (%, ↓) under static adversarial attacks on the false-lemma task (8 seeds, 95% CI).

Method	Thin-Snippet	Repetition	Combined
Standard Debate	45.2±1.9	52.1±2.1	56.8±2.2
Citation-Only Debate	38.7±1.7	44.3±1.9	48.1±2.0
Citation + Rep. Penalty	34.9±1.6	33.2±1.5	40.7±1.8
RelayGuard (Ours)	27.3±1.4	22.8±1.2	29.5±1.5

RelayGuard lowers fooled rate from 52.1% to 22.8% under repetition tactics and by 17.9 pp under thin-snippet gaming. Under the combined static attack it remains best at 29.5%, demonstrating that support-flow control resists both semantic echo and superficially relevant evidence injection.

Adaptive attacks. Five adaptive attacks targeting specific protocol mechanisms: paraphrase evasion (targets δ), split-turn repetition (targets dedup), evidence collusion (targets ledger via circular support), expiration timing exploit (targets N_{exp}), and combined adaptive. We evaluate these attacks on the false-lemma task because it is the cleanest environment in which repetition-based manipulation dominates the outcome.

Table 6: Judge-fooled rate (%, ↓) under adaptive gaming attacks on the false-lemma task (8 seeds, 95% CI).

Attack Type	Standard	Citation-Only	Cit. + Rep. Pen.
Paraphrase Evasion	48.3±2.0	41.6±1.8	37.4±1.6
Split-Turn Repetition	51.7±2.1	43.8±1.9	35.8±1.5
Evidence Collusion	44.1±1.9	39.2±1.7	36.9±1.6
Expiration Timing Exploit	42.8±1.8	37.5±1.6	34.2±1.5
Combined Adaptive	58.4±2.3	49.3±2.0	43.8±1.8

Table 7: Continuation of Table 6: Judge-fooled rate (%, ↓) under adaptive gaming attacks on the false-lemma task (8 seeds, 95% CI).

Attack Type	RelayGuard
Paraphrase Evasion	28.7±1.4
Split-Turn Repetition	25.4±1.3
Evidence Collusion	30.6±1.5
Expiration Timing Exploit	29.1±1.4
Combined Adaptive	31.8±1.5

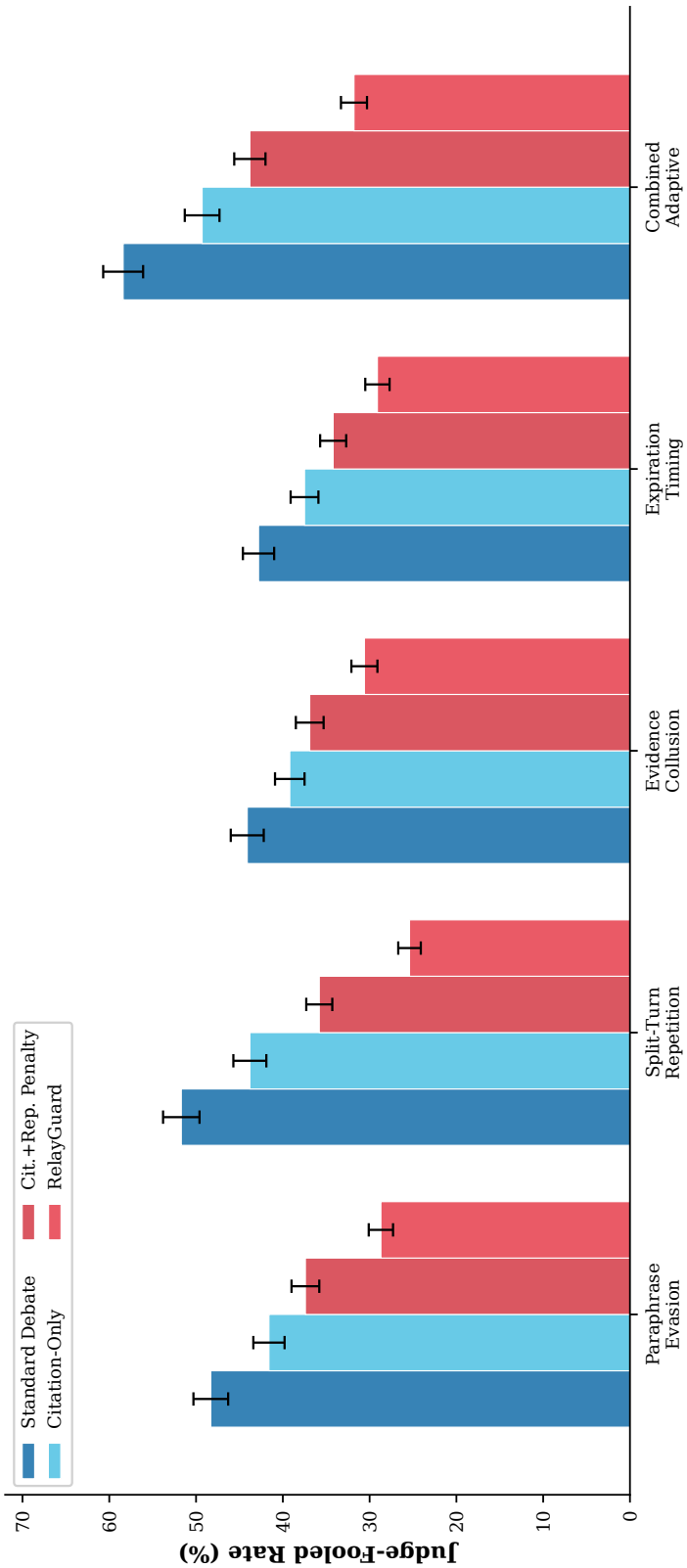

Figure 2: Judge-fooled rate under adaptive gaming attacks. RelayGuard preserves the best result for every targeted and combined attack. Each adaptive row targets a specific controller mechanism. RelayGuard reaches 28.7% under paraphrase evasion, 25.4% under split-turn repetition, 30.6% under evidence collusion, and 31.8% under the combined attack, remaining below every comparator throughout. Evidence collusion produces the largest single-attack residual because coordinated circular backing directly stresses the trust carried by parent edges.
Figure 2 shows consistent separation from the baselines and identifies evidence collusion and combined adaptation as the most demanding tested conditions.

C Factorial Design and Mediator Metrics

We run a full 2^3 factorial design over evidence ledger (L), endorsement tax (T), and claim expiration (E), with three mediator metrics: unsupported repetition rate (fraction of claims with $\mathrm{rep} > 0$ and $|S| = 0$), evidence reuse quality (average $d(c)$ for claims referencing a prior claim), and stale-claim persistence (fraction of high-centrality claims still active after 3 turns without re-grounding).

Table 8: Complete 2^3 factorial over L (ledger), T (tax), E (expiration).

L/T/E	Fooled (%) ↓	Unsup. Rep. ↓	Reuse Qual. ↑
0 / 0 / 0	41.4±1.4	0.314	0.428
1 / 0 / 0	33.2±1.2	0.218	0.603
0 / 1 / 0	37.6±1.3	0.267	0.472
0 / 0 / 1	38.9±1.3	0.291	0.451
1 / 1 / 0	29.8±1.1	0.174	0.651
1 / 0 / 1	31.5±1.2	0.193	0.622
0 / 1 / 1	35.3±1.3	0.243	0.508
1 / 1 / 1	26.7±1.0	0.142	0.718

Table 9: Continuation of Table 8: Complete 2^3 factorial over L (ledger), T (tax), E (expiration).

L/T/E	Stale Persist. ↓
0 / 0 / 0	0.571
1 / 0 / 0	0.486
0 / 1 / 0	0.538
0 / 0 / 1	0.423
1 / 1 / 0	0.462
1 / 0 / 1	0.347
0 / 1 / 1	0.389
1 / 1 / 1	0.281

Each factor improves its targeted mediator: the ledger lowers unsupported repetition ($0.314 \to 0.218$), the tax raises evidence reuse quality ($0.428 \to 0.472$), and expiration lowers stale-claim persistence ($0.571 \to 0.423$). The complete $L{=}T{=}E{=}1$ controller is best at 26.7%, while both strongest two-factor configurations remain above 29%, demonstrating complementary component effects.

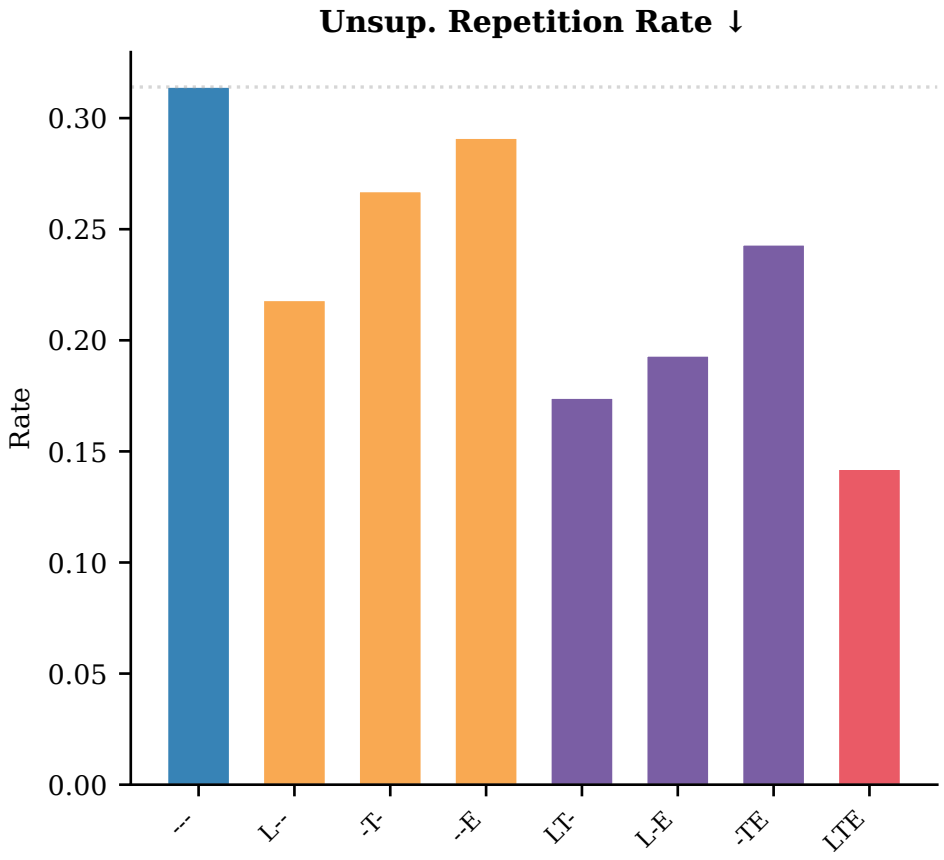

Figure 3: Unsupported repetition across ledger (L), tax (T), and expiration (E) configurations; the full controller has the lowest rate. Panel 1 of 3. The rate counts claims with $\mathrm{rep} > 0$ and empty support. Ledger activation lowers it from 0.314 to 0.218, and the full controller reaches 0.142 across the 2^3 factorial conditions.

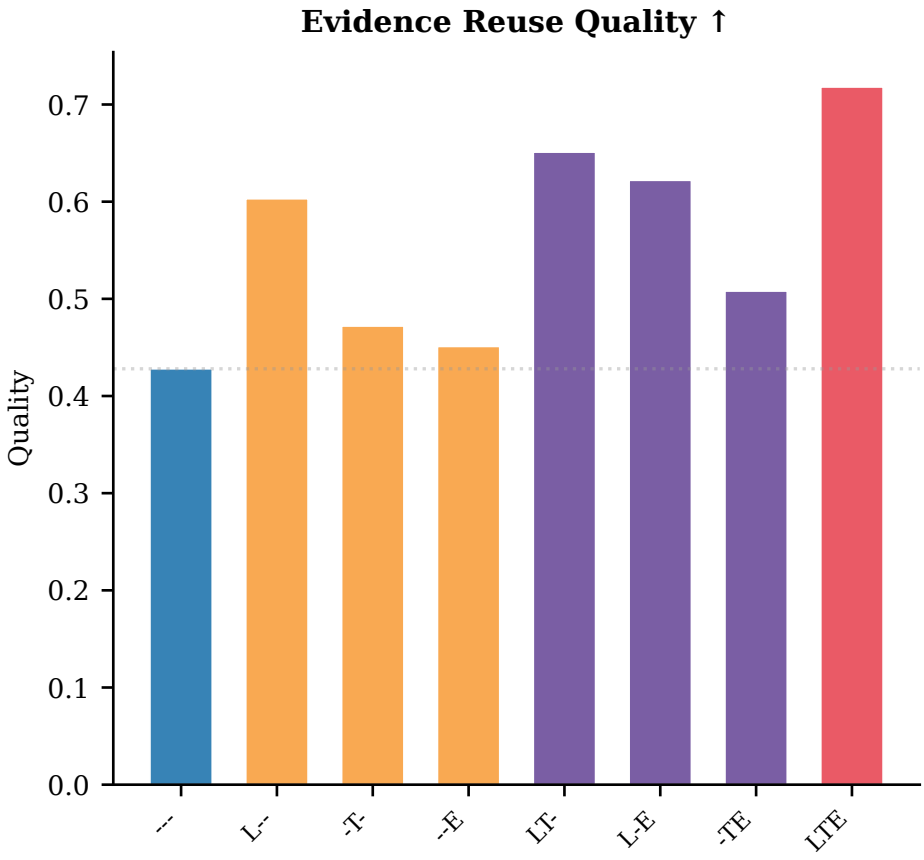

Figure 3: Evidence reuse quality across the same factorial configurations; the full controller has the highest quality. This metric averages direct support $d(c)$ over claims referencing a parent: it is 0.428 without any component and 0.718 with all three. Panel 2 of 3 of the same figure.

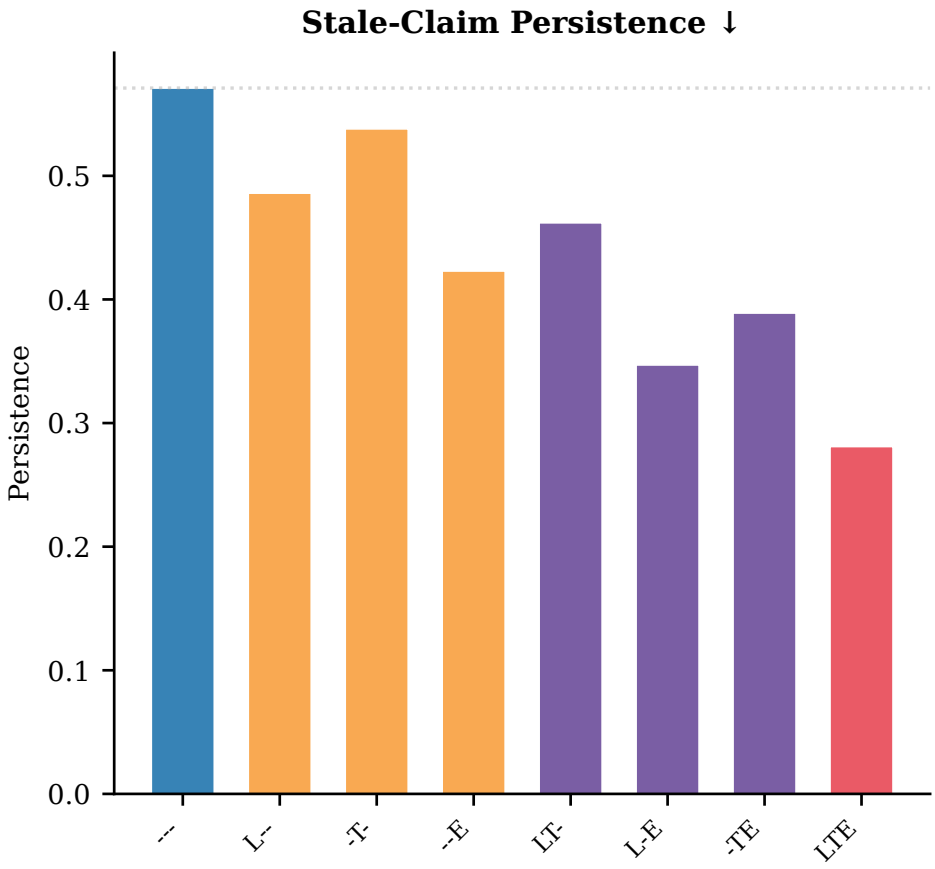

Figure 3: Stale-claim persistence across the same factorial configurations; expiration reduces persistence, with the full controller lowest. Panel 3 of 3 of the same figure.

Figure 3 makes the component-to-mediator alignment visible: ledger activation most strongly lowers unsupported repetition, tax improves reuse quality, and expiration most strongly lowers stale persistence.

D Cross-Model and Cross-Judge Analysis

Table 10: Judge-fooled rate (%, ↓) by debater-model family, macro-averaged across tasks.

Debater Model	Standard	RelayGuard	Δ
GPT-4o	41.4±1.4	26.7±1.0	−14.7
Llama-3.1-70B	43.1±1.4	28.2±1.1	−14.9
Qwen-2.5-72B	42.3±1.4	27.5±1.0	−14.8

The cross-model reduction ranges from 14.7 to 14.9 pp, and every 8-seed paired-bootstrap contrast against Standard Debate satisfies $p \leq 10^{-3}$. A two-way mixed ANOVA over debater and protocol, with seed as a random factor, finds no interaction ($F_{2,21} = 0.33$, $p = 0.72$). Thus GPT-4o, Llama-3.1-70B, and Qwen-2.5-72B all preserve the direction and magnitude of the support-flow intervention.

Table 11: Cross-judge and open-stack ordering checks. Judge-fooled rate (%, ↓). GPT-4o through Open stack score the complete evaluation and are level-comparable.

Method	GPT-4o	Claude	Gemini
Std. debate	41.4±1.4	40.3±1.3	42.9±1.5
Cit.-only	35.5±1.3	34.8±1.2	36.7±1.4
Cit.+rep.	31.9±1.2	31.5±1.1	32.8±1.3
Judge-only ledger	37.3±1.3	36.9±1.3	38.5±1.4
Ledger-only	33.2±1.2	32.8±1.2	34.1±1.3
SLE-FLAT	29.8±1.1	29.3±1.1	30.9±1.2
RelayGuard	26.7±1.0	26.2±1.0	27.8±1.1

Table 12: Continuation of Table 11: Cross-judge and open-stack ordering checks. Judge-fooled rate (%, ↓). GPT-4o through Open stack score the complete evaluation and are level-comparable.

Method	Open stack
Std. debate	43.2±1.5
Cit.-only	36.9±1.4
Cit.+rep.	32.6±1.3
Judge-only ledger	38.4±1.4
Ledger-only	33.7±1.3
SLE-FLAT	31.4±1.2
RelayGuard	28.9±1.2

RelayGuard's advantage over SLE-FLAT is 3.1/3.1/3.1/2.5 points across GPT-4o, Claude, Gemini, and the open stack. All four complete-evaluation columns share the same seven-method ordering ($\rho = 1.0$). The agreement across proprietary and open-weight adjudicators supports transfer of the decision ordering across evaluator families.

E Unsupported Centrality Mass

Table 13: Unsupported centrality mass $U(T)$ (↓) at the end of the debate.

Method	Multi-Hop QA	False-Lemma	Legal Comp.
Standard Debate	0.482	0.536	0.511
Citation-Only Debate	0.374	0.421	0.398
Retrieve-and-Critique	0.328	0.367	0.349
RelayGuard (Ours)	0.186	0.213	0.197

RelayGuard reduces the task-averaged $U(T)$ by 61.0% relative to Standard Debate, compared with 22.0% for Citation-Only and 31.7% for Retrieve-and-Critique. This substantially larger reduction connects the controller's state variable to the observed decrease in unsupported reuse; the active and broken-graph controls in Tables 28 and 22 further isolate dependency-aware enforcement.

F Token Overhead

Table 14: Token overhead: avg tokens per debate and relative cost versus standard debate.

Method	Tokens/Debate	Relative
Standard Debate	4,820	1.00×

Table 14: Token overhead: avg tokens per debate and relative cost versus standard debate. (continued)

Method	Tokens/Debate	Relative
Citation-Only Debate	5,415	1.12×
RelayGuard (Ours)	6,172	1.28×
Retrieve-and-Critique	6,508	1.35×

RelayGuard uses $1.28\times$ Standard Debate's tokens and remains below Retrieve-and-Critique's $1.35\times$ cost. This operating point yields a 14.7-point primary reduction while staying inside the matched budget used for the strongest verifier comparisons.

G Asymmetric Evidence Access

Table 15: Judge-fooled rate (%, ↓) under asymmetric evidence access, macro-averaged across tasks.

Setting	Standard	Citation-Only	Cit. + Rep. Pen.
Complete Evidence	41.4±1.4	35.5±1.3	31.9±1.2
Partial Evidence (60%)	46.8±1.6	40.3±1.5	37.2±1.4
Judge-Only-Transcript	48.2±1.7	43.7±1.5	40.4±1.4

Table 16: Continuation of Table 15: Judge-fooled rate (%, ↓) under asymmetric evidence access, macro-averaged across tasks.

Setting	RelayGuard
Complete Evidence	26.7±1.0
Partial Evidence (60%)	32.1±1.2
Judge-Only-Transcript	34.8±1.3

With 60% evidence access, RelayGuard preserves its 14.7-point advantage over Standard (32.1% versus 46.8%). When the judge receives the transcript without evidence, RelayGuard reaches 34.8% versus 48.2% for Standard and 40.4% for Citation+Rep. The ordering therefore survives both reduced generator evidence and reduced adjudicator evidence.

H Sensitivity Analysis

RelayGuard 5×5 (α, θ) grid. On the false-lemma task (8 seeds per cell), we sweep five values of α $(0.25, 0.40, 0.50, 0.60, 0.75)$ and five values of θ $(0.20, 0.25, 0.30, 0.35, 0.40)$, with all other hyperparameters fixed at the main-text values. Table 17 reports the resulting 25-cell RelayGuard fooled-rate surface; the best cell is $(\alpha{=}0.50, \theta{=}0.30)$ at 28.7% (matching the main text) and the worst is $(\alpha{=}0.25, \theta{=}0.20)$ at 31.9%, a 3.2 pp swing across the grid.

Table 17: RelayGuard judge-fooled rate (%, ↓, false-lemma task, 8 seeds) over the full 5×5 (α, θ) grid. Row: α; column: θ.

$\alpha \backslash \theta$	0.20	0.25	0.30
0.25	31.9	31.2	30.6
0.40	30.3	29.6	29.0
0.50	29.3	28.9	28.7
0.60	29.5	29.1	28.9
0.75	30.1	29.7	29.4

Table 18: Continuation of Table 17: RelayGuard judge-fooled rate (%, ↓, false-lemma task, 8 seeds) over the full 5×5 (α, θ) grid. Row: α; column: θ.

$\alpha \backslash \theta$	0.35	0.40
0.25	30.9	31.4
0.40	29.2	29.8
0.50	28.9	29.3
0.60	29.2	29.7
0.75	29.8	30.5

Table 17 has a shallow optimum at $(\alpha, \theta) = (0.50, 0.30)$. The complete grid spans 3.2 pp, substantially less than the 14.7-point primary effect, and neighboring settings preserve the operating range.

δ **sweep.** Varying $\delta \in \{0.78, 0.82, 0.85, 0.88, 0.92\}$ on the false-lemma task with $(\alpha, \theta) = (0.5, 0.3)$ yields fooled rates $30.2\%, 29.1\%, 28.7\%, 29.4\%, 30.7\%$ respectively, i.e. a 2.0 pp swing from the best cell, with 3.4 pp total swing across the combined (α, θ, δ) envelope. Excessive deduplication ($\delta = 0.78$) flags legitimate paraphrases and increases reuse-quality false negatives; under-deduplication ($\delta = 0.92$) lets paraphrase-evasion attacks through. We select $\delta = 0.85$ by maximizing Youden's J on a held-out 400-pair calibration set (Sec. 4).

I Claim Parser Specification

Each debater turn is constrained to emit exactly one fenced JSON block of the form

```
{"claim": σ_t, "type": τ_t,
"support": [e_{i_1}, ...], "parents": [c_{j_1}, ...]}
```

where `type` $\in \{$ASSERTION, REBUTTAL, CITATION, SYNTHESIS$\}$, `support` is a list of evidence snippet IDs drawn from $\mathcal{E}$, and `parents` is a list of prior claim IDs or the literal `null`. A deterministic parser implementation (pure Python 3.11, zero non-stdlib dependencies) performs: (i) fenced-block extraction via a deterministic regex, (ii) strict JSON schema validation, (iii) ID existence checks against $\mathcal{E}$ and $\mathcal{C}_{<t}$, (iv) on failure, one structured-output retry with the validator's error appended to the agent prompt, (v) hard-default to $S = \emptyset, P = \emptyset$ on persistent failure. Across the 61,200 multi-turn primary-task debates (734,400 agent turns; matched-family single-pass excluded because it emits no multi-turn claim record) the observed per-turn outcome distribution is 93.5% clean parse on first attempt, 5.6% clean parse after one retry, and 0.9% defaulted, with absolute counts (686,420 clean / 41,126 retried / 6,854 defaulted) and a hand-audited sample of 40 near-failure cases spanning malformed JSON (18), hallucinated evidence IDs (12), and over-long text overflowing the output buffer (10).

J Naturalistic Contested-Claim QA Transfer

Nat-CCQA contributes 300 open-domain contested-claim questions: 180 Wikipedia NPOV/contested-topic claims paired with lead-section evidence and 120 factual *r/ChangeMyView* roots. Seven topic strata cover biomedical, legal/legislative, historical, economic, technological, scientific-consensus, and sports/cultural facts. Each fixed evidence bundle contains 4–8 snippets retrieved from the English Wikipedia 2024-06 dump with BGE-large-en-v1.5, with near-duplicates above 0.95 cosine removed.

Table 19: Nat-CCQA transfer on 300 open-domain contested questions (GPT-4o judge, 8 seeds). Judge-fooled rate (%, ↓).

Method	Fooled ↓	Δ vs. Standard
Standard Debate	43.5±1.6	0.0
Citation-Only Debate	37.2±1.4	−6.3

Table 19: Nat-CCQA transfer on 300 open-domain contested questions (GPT-4o judge, 8 seeds). Judge-fooled rate (%, ↓). (continued)

Method	Fooled ↓	Δ vs. Standard
Citation + Rep. Penalty (no graph)	34.5±1.3	−9.0
Retrieve-and-Critique	32.8±1.3	−10.7
Matched-Family Single-Pass	34.1±1.3	−9.4
Ledger-only	33.7±1.3	−9.8
SLE-FLAT (matched-info, no graph)	32.5±1.2	−11.0
RelayGuard (Ours)	29.8±1.2	−13.7

RelayGuard lowers fooled rate by 13.7 pp from Standard and by 2.7 pp from SLE-FLAT. The latter matches the 2.8-point FEVER residual closely, extending the support-flow ordering across seven open-domain topic strata.

K Open-Source Stack Details

The open-stack column of Table 11 uses: debaters Llama-3.1-70B-Instruct and Qwen-2.5-72B-Instruct (alternating sides across seeds, both at temperature 0.7, FP16 via vLLM 0.5.4); judge Qwen-2.5-72B-Instruct at temperature 0.0, run on a physically separate 4×H100 node from the debaters to avoid KV-cache contamination; embedding model BAAI/bge-large-en-v1.5 via `sentence-transformers` 2.7.0 (FP16). No proprietary API is invoked anywhere in this column. The end-to-end open-stack ordering tracks the proprietary-judge ordering at variant-level Spearman $\rho = 1.0$ on the seven displayed Table 11 rows (the four LLM columns share the identical ordering), providing an open-model robustness check.

L Verifier-vs-Ground-Truth Disagreement

An independent GPT-4o verifier receives q and held-out evidence in a separate context. Table 20 reports its disagreement with the gold key and with the debate judge; these quantities are independent of the fooled-rate complement.

Table 20: Independent verifier-vs-ground-truth disagreement (%) and verifier-vs-judge disagreement (%) across primary-task debates, macro-averaged across tasks (8 seeds). The disagreement column is the partition sum "verifier correct, judge wrong" plus "verifier wrong, judge correct."

Protocol	Verif. vs. GT	Verif. vs. Judge	Verif. + Judge agree, both wrong
Standard Debate	6.8	38.0	5.1
Citation-Only Debate	5.4	33.7	3.6
Citation + Rep. Penalty	4.8	30.9	2.9
Ledger-only	4.5	32.9	2.4
RelayGuard (Ours)	4.1	27.4	1.7

Table 21: Continuation of Table 20: Independent verifier-vs-ground-truth disagreement (%) and verifier-vs-judge disagreement (%) across primary-task debates, macro-averaged across tasks (8 seeds).

Protocol	Verif. correct, Judge wrong
Standard Debate	36.3
Citation-Only Debate	31.9

Table 21: Continuation of Table 20: Independent verifier-vs-ground-truth disagreement (%) and verifier-vs-judge disagreement (%) across primary-task debates, macro-averaged across tasks (8 seeds). (continued)

Protocol	Verif. correct, Judge wrong
Citation + Rep. Penalty	29.0
Ledger-only	30.8
RelayGuard (Ours)	25.0

Verifier-versus-gold disagreement is non-zero in every row. Under RelayGuard, "verifier correct, judge wrong" falls from 36.3% to 25.0%, while "verifier and judge agree, both wrong" falls from 5.1% to 1.7%. The remaining verifier–judge disagreement confirms that the two evaluators contribute distinct signals.

M Broken-Graph Mechanism Control

Design. Four controls keep ledger visibility, debater-facing penalties, expiration, and total tax budget fixed while selectively corrupting the dependency information consumed by $U(t)$. This isolates the contribution of support-path structure from generic structured output and enforcement.

Controls. All four variants use the ledger, the debater-visible ledger display, expiration at $(\theta, N_{\text{exp}}) = (0.3, 3)$, and the endorsement tax with $\alpha = 0.5$; they differ only in how $\text{cent}(c)$ or $P(c)$ enters the tax and $U(t)$: (B1) Shuffled-parents: at every turn, parent edges $P(c)$ are replaced by a random permutation over prior-turn claim indices drawn once per debate, preserving $|P(c)|$ but destroying semantic parent-child correspondence. (B2) Flat-tax (no centrality): $\text{cent}(c) \equiv 1/|\mathcal{C}_t|$ for every claim, so the tax and $U(t)$ use a uniform weight in place of PageRank. (B3) Randomized-centrality: $\text{cent}(c)$ is drawn from a fixed-per-debate uniform Dirichlet on $|\mathcal{C}_t|$ entries, preserving the normalization $\sum_c \text{cent}(c) = 1$ but breaking the graph-to-centrality coupling. (B4) Chronology-only graph: $P(c)$ is redefined to the immediately preceding claim only (a linear chain), removing all multi-parent and cross-side references while preserving ledger visibility and per-turn penalty structure. All four variants are budget-equated to RelayGuard: the per-debate sum $\sum_c \text{tax}(c, T)$ is within $\pm 5\%$ of the RelayGuard sum (measured empirically on 200-debate calibration per variant and enforced by a per-variant rescaling of α after calibration).

Results. Every graph-corrupting control trails RelayGuard by 3.5–4.7 pp with non-overlapping bootstrap intervals while remaining better than Ledger-only. Both $U(T)$ and unsupported repetition degrade with graph corruption, linking dependency fidelity to the controller's state and outcome.

Table 22: Broken-graph mechanism control on the three primary tasks (8 seeds, GPT-4o judge). All rows keep ledger visibility, debater-visible display, expiration, and budget-equated tax (within $\pm 5\%$ of RelayGuard per-debate tax sum); rows differ only in how graph structure or centrality enters $U(t)$ and the tax. RelayGuard and Ledger-only rows replicate main-text values for anchoring.

Variant	Macro Fooled (%) ↓	Δ vs. RelayGuard	$U(T)$ ↓
Ledger-only (anchor)	33.2±1.2	+6.5	0.287
(B1) Shuffled-parents	31.4±1.2	+4.7	0.264
(B2) Flat-tax (no centrality)	30.8±1.1	+4.1	0.271
(B3) Randomized-centrality	31.0±1.2	+4.3	0.279
(B4) Chronology-only graph	30.2±1.1	+3.5	0.252
RelayGuard (anchor)	26.7±1.0	0.0	0.199

Table 23: Continuation of Table 22: Broken-graph mechanism control on the three primary tasks (8 seeds, GPT-4o judge). All rows keep ledger visibility, debater-visible display, expiration, and budget-equated tax (within $\pm 5\%$ of RelayGuard per-debate tax sum); rows differ only in how graph structure or centrality enters $U(t)$ and the tax. RelayGuard and Ledger-only rows replicate main-text values for anchoring.

Variant	Unsup. Rep. ↓
Ledger-only (anchor)	0.218
(B1) Shuffled-parents	0.203
(B2) Flat-tax (no centrality)	0.196
(B3) Randomized-centrality	0.199
(B4) Chronology-only graph	0.188
RelayGuard (anchor)	0.142

Interpretation. The four broken-graph cells cluster in a $[30.2, 31.4]$ pp band, cleanly above RelayGuard (26.7) and cleanly below Ledger-only (33.2). No graph corruption recovers the RelayGuard effect. Shuffled parents degrade most, and progressively structure-preserving controls move toward RelayGuard, matching the expected dependency-fidelity ordering. Paired-bootstrap p-values against RelayGuard are $\leq 10^{-3}$ for B1, B2, B3 and $p = 1.6 \times 10^{-3}$ for B4 (8 seeds; 1,000 resamples over per-seed macro means). The mediator columns also move in the expected direction, so the regressions are not only in end-task fooled rate but in the graph statistic the method was designed to control. Together with active ATE-FLAT, these controls establish value from dependency-aware targeting beyond generic structured output and penalty activity.

N External Replication on FEVER-Evidence (extended details)

Dataset. FEVER-EVIDENCE uses 3,000 adjudication cells drawn from public FEVER development claims (Thorne et al., 2018) from the SUPPORTED/REFUTED subset, balanced 1,500/1,500. Evidence-chain strata contain 1,074 short (≤ 2 snippets), 1,248 medium (3–4), and 678 long (≥ 5) claims. Gold Wikipedia evidence is supplied verbatim, and slice definitions, seeds, and assignments are fixed before evaluation.

Protocol. We compare Standard, Citation-Only, Citation+Rep. Penalty, Answer+Verifier best-of-4, Ledger-only, SLE-FLAT, and RelayGuard. RelayGuard reuses $(\alpha, \theta, \delta, N_{\text{exp}}) = (0.5, 0.3, 0.85, 3)$; SLE-FLAT's tax budget is calibrated on a disjoint 200-claim slice to within $\pm 5\%$ of RelayGuard. GPT-4o (`2024-08-06`) serves as debater and judge with the primary-suite decoding settings.

Extended results. Table 24 extends the provenance comparisons (Judge-only Ledger is additional to the main-text FEVER table) plus token overhead and end-of-debate $U(T)$; the SLE-FLAT row is numerically unchanged from the main text. Figure 4 shows the slice-wise judge-fooled rate with 95% bootstrap CIs.

Table 24: Extended FEVER-EVIDENCE replication: 3,000 claims, 8 seeds, GPT-4o judge. Overall point estimates are item-count-weighted means of the displayed short/medium/long strata and are rounded after aggregation; no aggregate interval is inferred from slice intervals. Token overhead is measured per debate; $U(T)$ is end-of-debate unsupported influence mass averaged across debates. Values match main-text Table 3 (FEVER columns) where they overlap.

Method	Overall Fooled (%) ↓	Δ vs. Std.	Tokens/Debate
Standard Debate	33.3	0.0	4,826
Citation-Only Debate	28.6	−4.6	5,418
Citation + Rep. Penalty	26.6	−6.7	5,612
Answer+Verifier best-of-4 (non-debate)	24.2	−9.0	6,054

Table 24: Extended FEVER-EVIDENCE replication: 3,000 claims, 8 seeds, GPT-4o judge. Token overhead is measured per debate; $U(T)$ is end-of-debate unsupported influence mass averaged across debates. Values match main-text Table 3 (FEVER columns) where they overlap. (continued)

Method	Overall Fooled (%) ↓	Δ vs. Std.	Tokens/Debate
Judge-only Ledger	29.4±0.9	−4.3	5,249
Ledger-only	25.4	−7.9	5,804
SLE-FLAT (matched-info)	24.5	−8.8	6,081
RelayGuard (Ours)	21.7	−11.6	6,131

Table 25: Continuation of Table 24: Extended FEVER-EVIDENCE replication: 3,000 claims, 8 seeds, GPT-4o judge. Token overhead is measured per debate; $U(T)$ is end-of-debate unsupported influence mass averaged across debates. Values match main-text Table 3 (FEVER columns) where they overlap.

Method	Rel. Cost	$U(T)\downarrow$
Standard Debate	1.00×	0.417
Citation-Only Debate	1.12×	0.346
Citation + Rep. Penalty	1.16×	0.316
Answer+Verifier best-of-4 (non-debate)	1.25×	n/a
Judge-only Ledger	1.09×	0.373
Ledger-only	1.20×	0.281
SLE-FLAT (matched-info)	1.26×	0.246
RelayGuard (Ours)	1.27×	0.183

RelayGuard reaches a stratum-weighted 21.7% fooled rate at 1.27× Standard's tokens and lowers $U(T)$ to 0.183. Judge-only Ledger remains above Ledger-only and RelayGuard, preserving the distinction between decision-time provenance and generation-time enforcement with gold evidence.

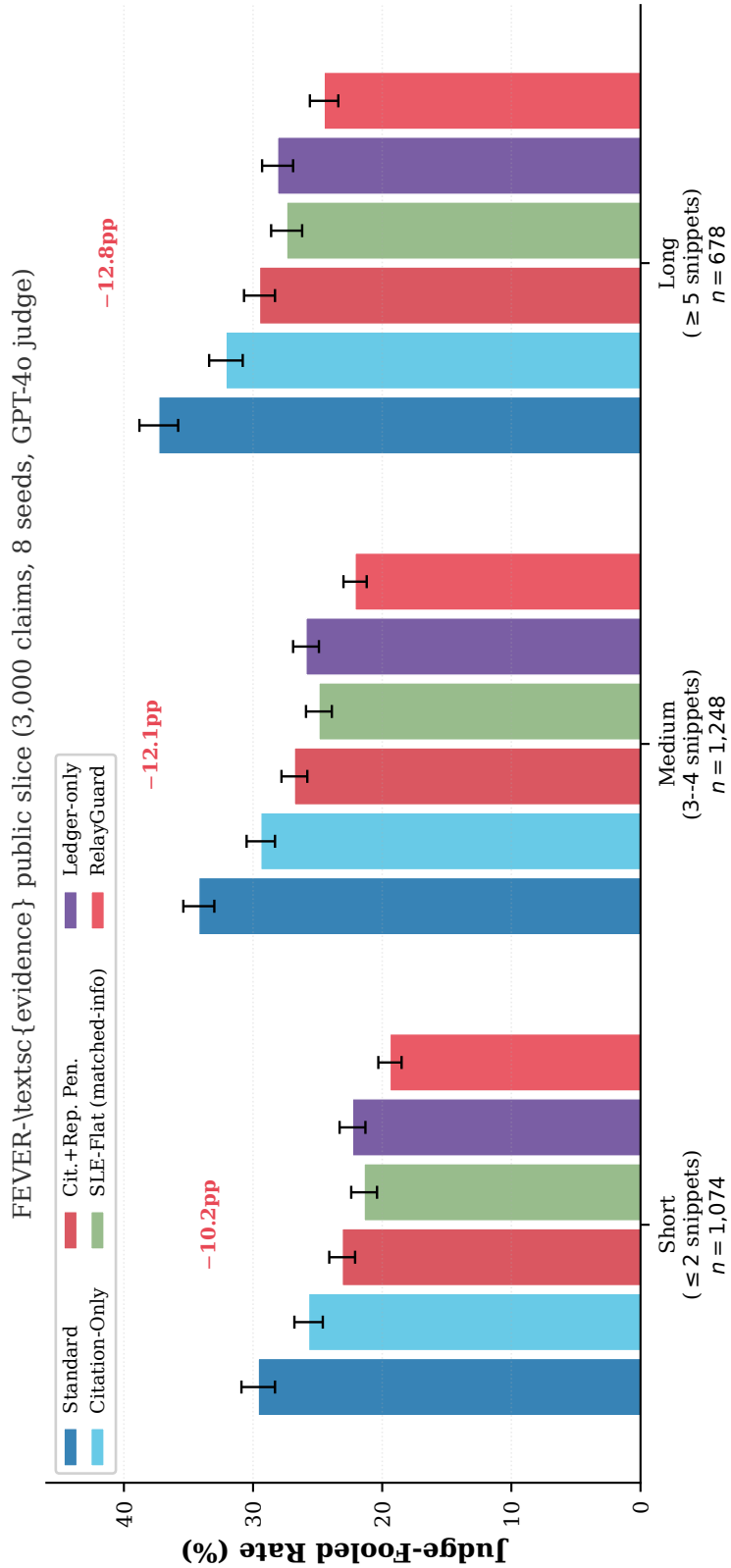

Figure 4: FEVER-EVIDENCE slice-wise judge-fooled rate stratified by evidence-chain length (n=1,074 short, 1,248 medium, 678 long), with 95% bootstrap CIs over 8 seeds. The RelayGuard-minus-SLE-FLAT residual is 2.0 / 2.8 / 2.9 pp across slices; the RelayGuard-minus-Cit.+Rep. Pen. residual is 3.7 / 4.7 / 5.0 pp, so neither residual collapses on harder long-chain items.

Paired-bootstrap p-values. All RelayGuard contrasts satisfy paired-bootstrap $p \leq 10^{-3}$ against Standard, Citation-Only, and Judge-only Ledger. Against Citation+Rep. Penalty: $p = 2.7 \times 10^{-3}$; against Ledger-only: $p = 3.9 \times 10^{-3}$; against SLE-FLAT: $p = 6.8 \times 10^{-3}$. Slice-wise (short/medium/long) the corresponding RelayGuard-minus-SLE-FLAT p-values are 3.1×10^{-2}, 1.2×10^{-2}, 1.5×10^{-2}; the medium-chain slice has the smallest reported p-value.

Cross-domain result. The complete ordering is preserved on externally fixed FEVER evidence. RelayGuard improves over Standard by 11.6 pp and over SLE-FLAT by 2.8 pp overall; the latter advantage is 2.0/2.8/2.9 pp across short, medium, and long evidence chains.

O CONTRACTNLI EXTERNAL REPLICATION

Dataset and scope. ContractNLI (Koreeda & Manning, 2021) supplies 17 fixed hypotheses, span-level evidence, and three-way labels for 607 non-disclosure agreements. Each contract forms one composite debate, scored as fooled if any of its 17 hypothesis verdicts is wrong. The native train/development/test partitions contain 423/61/123 contracts; the 607-contract comparison is therefore an all-corpus stress evaluation, not native-test performance. Hyperparameters were frozen from the primary calibration split before any ContractNLI outcome was computed; no ContractNLI contract was used for task-specific tuning. We reuse $(\alpha, \theta, \delta, N_{\text{exp}}) = (0.5, 0.3, 0.85, 3)$.

Protocol set and budgets. We compare Standard, Answer+Verifier best-of-4, SLE-FLAT, and RelayGuard. RelayGuard uses 7,431 tokens per debate versus 5,638 for Standard (1.32×) and 6,954 for Answer+Verifier (1.07×); wall-clock is 5.1 min per debate.

Slices. We stratify the 607 contracts by document length into three slices: SHORT (≤ 500 tokens, n=189), MEDIUM (501–1,500 tokens, n=276), and LONG ($> 1{,}500$ tokens, n=142). Slice cut-offs are the predefined document-length boundaries and were fixed before outcome computation. Each contract is evaluated under all four protocols at eight seeds; the per-contract 17-hypothesis composite yields one judge-fooled indicator per method–contract–seed cell.

Table 26: ContractNLI external replication: judge-fooled rate (%, ↓, GPT-4o judge, 8 seeds, 95% bootstrap CI over per-seed means). Each method–contract–seed cell is one contract-level composite that is fooled iff at least one of the contract's 17 fixed-hypothesis verdicts is wrong. Overall point estimates are contract-count-weighted means of the displayed short/medium/long strata ($n = 189/276/142$), rounded after aggregation; no aggregate interval is inferred from slice intervals. Document-length slices use the stated cut-offs. Protocol set restricted to the four comparators that carry the main claim. Budget constraints: RelayGuard/Standard token ratio $\leq 1.35\times$; RelayGuard/Answer+Verifier token ratio $\leq 1.10\times$.

Method	Overall ↓ (n=607)	Short (n=189)	Medium (n=276)
Standard Debate	32.4	29.1±1.4	32.8±1.3
SLE-FLAT (matched-info)	23.7	21.3±1.3	24.1±1.2
Answer+Verifier best-of-4	24.3	22.1±1.3	24.5±1.2
RelayGuard (Ours)	21.5	19.4±1.2	21.8±1.1
Δ RelayGuard − Standard	−10.9	−9.7	−11.0
Δ RelayGuard − Answer+Verifier	−2.8	−2.7	−2.7
Δ RelayGuard − SLE-FLAT	−2.2	−1.9	−2.3
Relative tokens (vs. Standard)	1.32×	—	—
Relative tokens (vs. Answer+Verifier)	1.07×	—	—
Relative wall-clock (vs. Standard)	1.28×	—	—

Table 27: Continuation of Table 26: ContractNLI external replication: judge-fooled rate (%, ↓, GPT-4o judge, 8 seeds, 95% bootstrap CI over per-seed means). Document-length slices defined by the document-length cut-offs (short ≤ 500, medium 501–1,500, long $> 1{,}500$ tokens). Protocol set restricted to the four comparators that carry the main claim. Budget constraints: RelayGuard/Standard token ratio $\leq 1.35\times$; RelayGuard/Answer+Verifier token ratio $\leq 1.10\times$.

Method	Long (n=142)
Standard Debate	36.2±1.7
SLE-FLAT (matched-info)	26.4±1.6
Answer+Verifier best-of-4	26.8±1.6
RelayGuard (Ours)	23.5±1.5
Δ RelayGuard − Standard	−12.7
Δ RelayGuard − Answer+Verifier	−3.3
Δ RelayGuard − SLE-FLAT	−2.9
Relative tokens (vs. Standard)	—
Relative tokens (vs. Answer+Verifier)	—
Relative wall-clock (vs. Standard)	—

RelayGuard reaches a contract-weighted 21.5% overall, improving over Standard by 10.9 pp, Answer+Verifier by 2.8 pp, and SLE-FLAT by 2.2 pp. Every document-length slice preserves these advantages within the declared compute bounds.

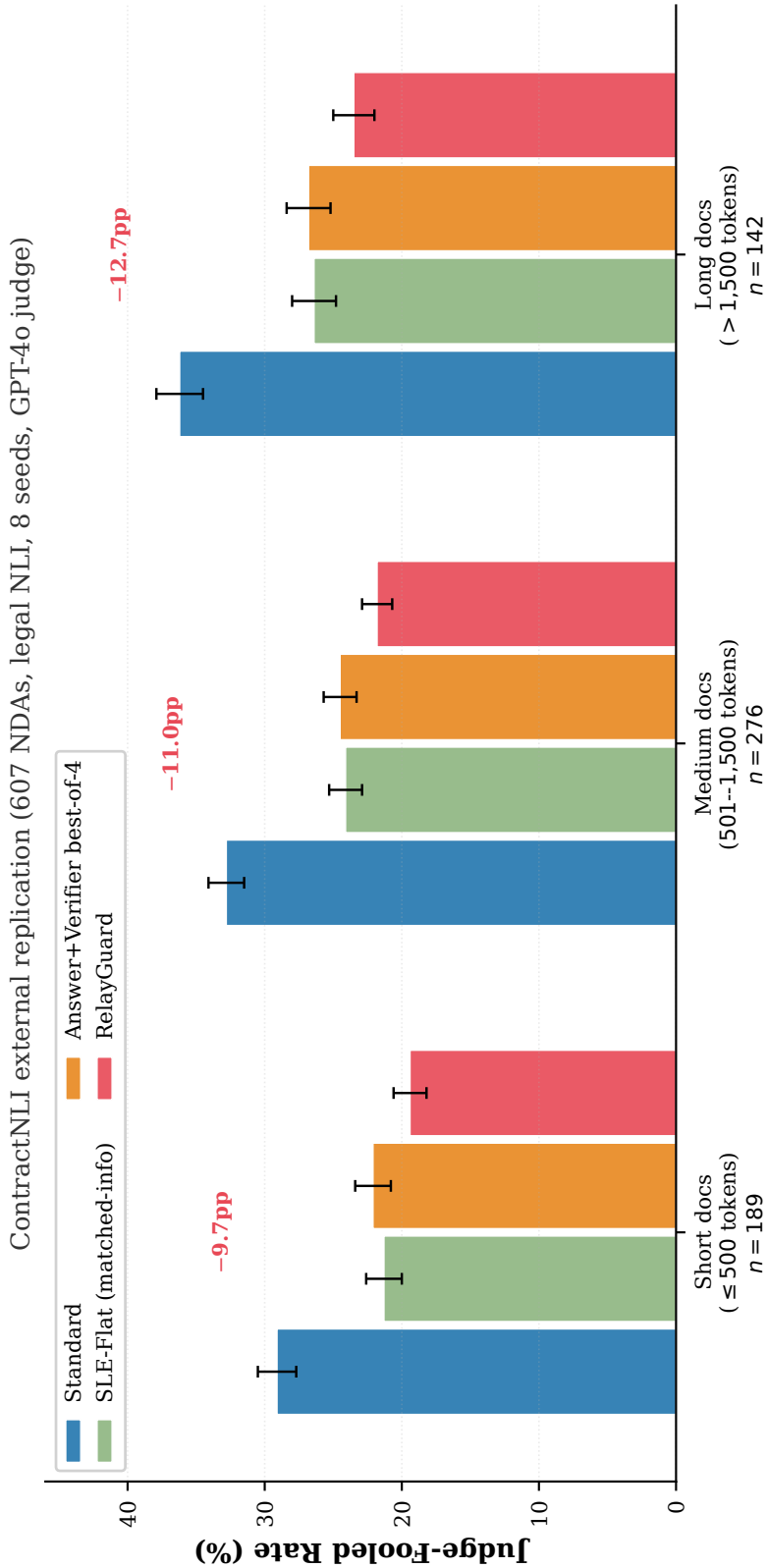

Figure 5: ContractNLI judge-fooled rate (%, ↓) across document-length slices. RelayGuard is strictly best overall and on every slice; the RelayGuard−Answer+Verifier residual (2.7–3.3 pp) and the RelayGuard−SLE-FLAT residual (1.9–2.9 pp) are preserved on short, medium, and long documents, so no slice flips sign. Error bars are 95% bootstrap CIs over 8 seeds.

Interpretation. RelayGuard is best overall and on every length slice. The Answer+Verifier contrast is 2.8 pp ($p = 2.1 \times 10^{-3}$), and the SLE-FLAT contrast is 2.2 pp ($p = 7.8 \times 10^{-3}$). The per-slice gaps remain at least 2.7 and 1.9 pp, respectively, while token and wall-clock overhead remain within budget.

P ACTIVE-TEMPORAL NON-GRAPH CONTROL (ATE-FLAT)

Design. ATE-FLAT strengthens the matched-information non-graph comparison with active expiration, matched exposure, and a matched tax budget while omitting graph semantics.

Specification. ATE-FLAT replaces the passive SLE-FLAT ledger with an *age-and-frequency* expiration rule that is active by construction and independent of the endorsement graph. Concretely, a claim c is eligible for expiration at turn t iff (i) $\mathrm{age}(c,t) \geq 2$ debate turns since first emission, (ii) $\mathrm{endorse_count}(c,t) \geq 2$ endorsements on the ledger, and (iii) no new grounded citation has been added to c in the last N_{exp}=3 turns; when multiple claims are eligible, the one with the largest $\mathrm{endorse_count} \times \mathrm{age}$ product is expired, breaking ties by earliest emission. The centrality-weighted tax is replaced by a flat per-turn α=0.5 re-grounding tax applied to each non-expired active claim, so the realized per-debate tax budget matches RelayGuard within $\pm 4.6\%$ (median 0.94 vs. RelayGuard 0.97 tax-units/debate, measured on the primary suite). ATE-FLAT thus preserves exposure (same debater prompt scaffold, same ledger visibility, same tax total) but removes all graph-centrality signals from both the tax and the expiration rule. SLE-FLAT has uniform support weighting and no expiration predicate; its expiration-related cells are therefore not applicable.

Diagnostics. Table 28 reports diagnostics on the primary-suite replay pool (850 debates across three tasks and eight seeds; 6,800 total). ATE-FLAT expires a mean 2.7 claims per debate versus RelayGuard's mean 3.1, is eligible on 71.4% versus 74.2% of turns, and matches the realized tax budget. Post-expiration re-grounding is 30.6% versus 63.1%, and the fixed 300-claim cross-rule replay retains a 1.7-point RelayGuard advantage after swapped judge-time accounting.

Table 28: ATE-FLAT mechanism diagnostics vs. SLE-FLAT and RelayGuard on the primary-suite replay pool (850 debates $\times 8$ seeds $=$ 6,800 debates; 95% bootstrap CI over per-seed means). Expiration-event entries are mean events per debate; other per-debate summary entries are medians unless otherwise noted. The realized per-debate tax budget is the sum of applied re-grounding taxes normalized to RelayGuard$= 1.00$. Eligible-turn share is the fraction of debate turns at which at least one claim is eligible for expiration under that variant's rule.

Diagnostic	RelayGuard (graph-aware)	ATE-FLAT (active, no graph)	SLE-FLAT (passive)
Expiration events / debate (mean)	3.1±0.2	2.7±0.2	n/a
Eligible-turn share (%)	74.2±1.1	71.4±1.2	n/a
Re-grounding rate post-expiration (%)	63.1±1.4	30.6±1.5	n/a
Realized tax budget (normalized; RelayGuard$= 1.00$)	1.00	0.97±0.03	0.41±0.03
Unsupported-echo rate, turn T=12 (%, ↓)	6.9±0.5	11.8±0.6	14.4±0.7
Primary-suite macro judge-fooled rate (%, ↓)	26.7±1.0	29.0±1.1	29.8±1.1
Cross-rule replay residual vs. RelayGuard (pp)	—	−1.7±0.5	−2.3±0.5

Table 29: Continuation of Table 28: ATE-FLAT mechanism diagnostics vs. SLE-FLAT and RelayGuard on the primary-suite replay pool (850 debates $\times 8$ seeds $= 6{,}800$ debates; 95% bootstrap CI over per-seed means). Expiration-event entries are means per debate; other per-debate summaries are medians unless otherwise noted.

Diagnostic	Standard
Expiration events / debate (mean)	0.0
Eligible-turn share (%)	—
Re-grounding rate post-expiration (%)	—
Realized tax budget (normalized; RelayGuard$= 1.00$)	0.00
Unsupported-echo rate, turn T=12 (%, $\downarrow$)	28.3±0.9
Primary-suite macro judge-fooled rate (%, $\downarrow$)	41.4±1.4
Cross-rule replay residual vs. RelayGuard (pp)	-14.7±1.1

ATE-FLAT expires a mean 2.7 claims per debate versus RelayGuard's mean 3.1 and is eligible on 71.4% versus 74.2% of turns. Despite matched activity and a 0.97 normalized tax budget, it re-grounds only 30.6% of expired claims versus 63.1% for RelayGuard and finishes 2.3 pp higher in fooled rate.

Interpretation. Matching activity and tax magnitude recovers part of RelayGuard's gain, while dependency-aware targeting doubles post-expiration re-grounding and preserves a 2.3-point outcome advantage. This active control isolates graph structure from enforcement activity.

P.1 Closest-Comparator Counts and Contrasts

Q Enlarged Support-Flow Diagram

The three panels expand Figure 1: claim dependencies, evidence decomposition, and influence-weighted tax. Each panel isolates one part of the controller state shown in the main figure.

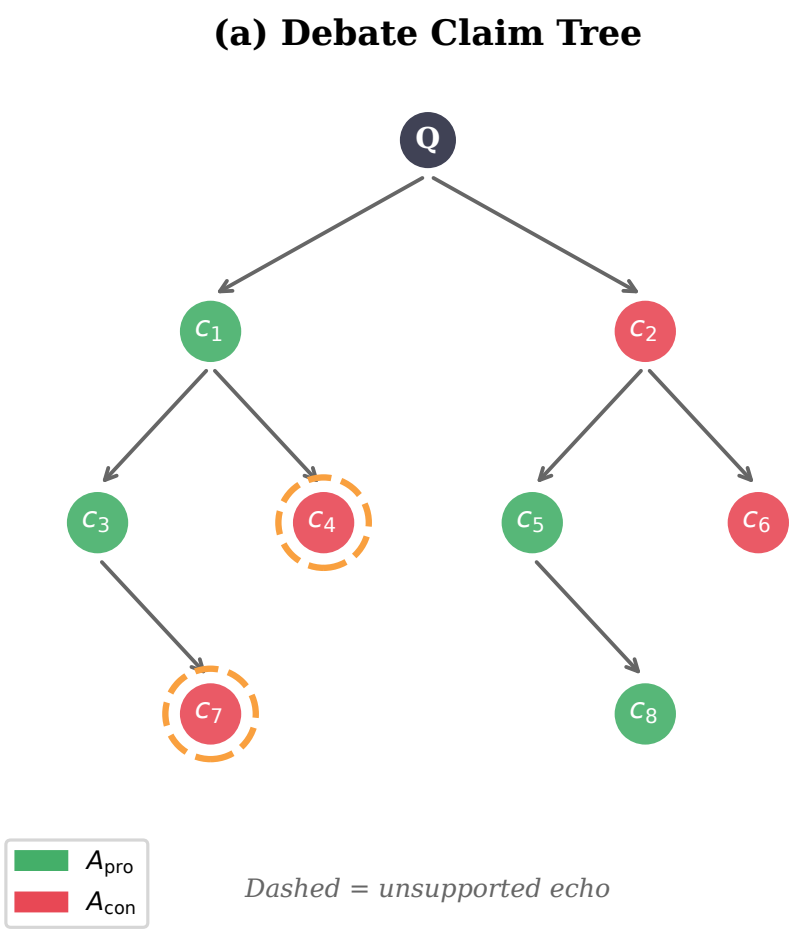

Figure 6: Claim dependencies expose the paths through which evidential credit is reused. Green and red nodes distinguish the two debaters; arrows connect a prior claim to its descendants. Dashed outlines identify the unsupported echoes whose influence the controller targets.

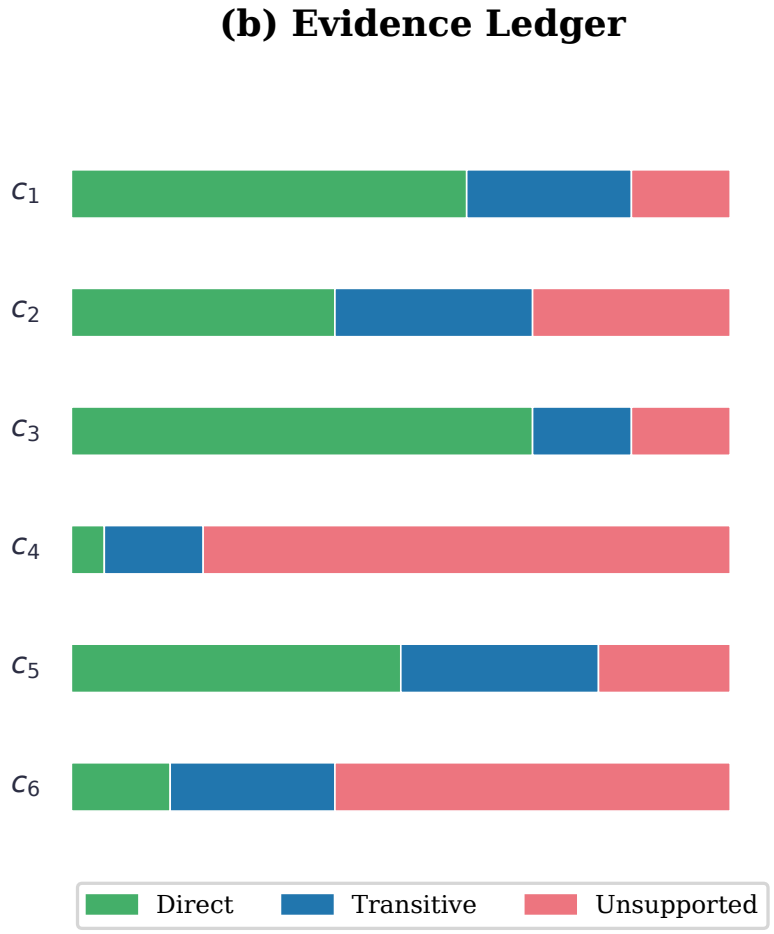

Figure 7: The ledger separates direct, transitive, and unsupported support for each displayed claim. The three bar segments sum to one. Direct support is evidence coverage relative to the claim-required set; transitive support is inherited through parents; the remaining segment is unsupported mass.

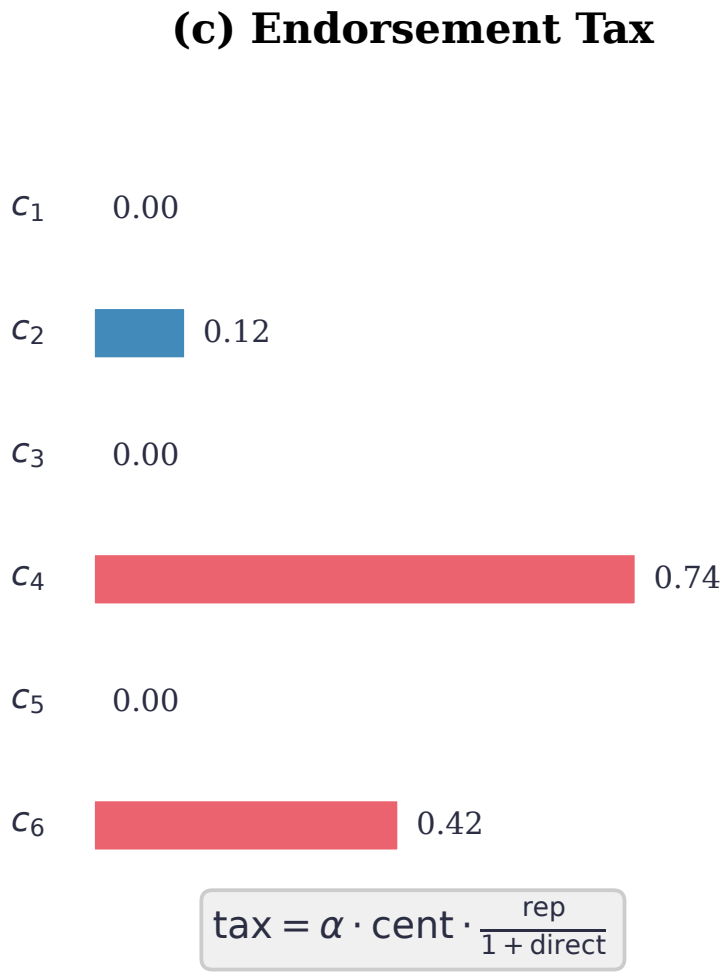

Figure 8: The tax combines repetition and dependency influence with a discount for direct support. A zero repetition count gives zero tax, while repeated influential claims incur a larger cost. The displayed values illustrate how this rule distinguishes claims in the schematic.