BLAMESHIFT: BLAME INTEGRITY OF MULTI-AGENT FAILURE ATTRIBUTION UNDER OWNERSHIP-CONSTRAINED NARRATIVE EDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent failure attribution relies on mixed logs combining trusted execution events with agent-authored messages. An agent responsible for a failure can exploit this interface by rewriting its own narrative to shift blame onto an innocent peer while execution records remain fixed. We define *blame integrity* as the invariance of an audit interface–judge mechanism to ownership-constrained narrative edits, precluding profitable deviations for blame-avoiding agents. On multi-agent benchmarks, ownership-legal accusations reliably steer flat LLM judges toward preselected victims over matched controls, an effect driven by peer accusations rather than text length, name priming, or factual fabrication. Vulnerability varies across judges, belonging to the audit mechanism rather than the text alone. Theoretically, evidence-independent style filters are either bypassable or own-blame inert, and empirically, filter-aware rewrites circumvent them. While narrative deletion trades off clean accuracy and citation verification limits steering, offline committees fail to Pareto-dominate their strongest constituent.

1 INTRODUCTION

When multi-agent tool-using systems fail, auditors inspect post-execution records to determine which agent and which step were decisive (Zhang et al., 2025; Sun et al., 2026). Those records often mix trusted execution events with messages and rationales authored by the agents under review: a *mixed audit view*. If an attributor treats writable narratives as evidence, a responsible agent can attempt *targeted blame shifting*—redirecting a fixed failure onto a preselected innocent peer by editing only its normally writable audit text, while actions, tool receipts, environment state, and the responsibility reference stay fixed. We call invariance to such edits *blame integrity*: a property of the audit interface φ together with the judge J, which implies strategyproofness for any responsible agent that prefers not to be blamed.

Who&When formalized automated failure attribution and showed that even strong models remain far from reliable step-level accuracy (Zhang et al., 2025). Later work improves clean diagnosis through two-stage search (Sun et al., 2026), learned tracers (Zhang et al., 2026), causal graphs (Wang et al., 2026), fuller observability (Chen et al., 2026), and intervention checks (Ma et al., 2026; Shah, 2026). They evaluate attribution on fixed logs. Chain-of-thought rewrites can flip outcomes with actions held fixed (Khalifa et al., 2026), and traces can relay adversarial content into monitors (Isbarov & Kantarcioglu, 2026), but neither asks whether a known failure can be reassigned to a chosen innocent agent under a field-ownership write boundary, or which interfaces block that reassignment.

Table 4 (Appendix B) shows two Who&When Hand-Crafted (HC) examples. After an owned end-of-trace F1 rationale that names Orchestrator, a flat All-at-Once judge flips from the true culprit (WebSurfer) to the named victim and echoes the accusation; naming Assistant instead misses and lands on Orchestrator. At scale, under ownership and immutability checks and matched noise (same victim, writer, placement, and judge), end-of-trace F1 rationales induce large targeted lifts for Qwen-Max All-at-Once on Who&When and TraceElephant. GPT, Kimi, Claude, and Gemini-2.5-Pro show substantially smaller lifts on the same attacked text. Susceptibility is thus a property of the pair (φ, J) rather than of the attack, which is what makes it an interface design variable: the same mixed audit view is resistant to this tested edit under one attributor and steerable under another.

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.

Contributions.

1. We formalize blame integrity as invariance of (φ, J) under field-ownership edits, relate it to strategyproofness (Remark 1), and characterize interface classes: filters that ignore trusted evidence are bypassable or degenerate (Proposition 1), citation checks leave selective disclosure (Proposition 2), and submission-independent disclosure gives judge-independent integrity (Corollary 2).
2. On Who&When and TraceElephant, ownership-legal named F1 edits induce large paired lifts for a flat All-at-Once judge; cross-family writers reproduce the effect under matched controls, and it survives removing the attacker's knowledge of $Z^\star$ and of the target.
3. On a cost–robustness frontier, lexical filtering fails under a filter-aware rewrite, narrative deletion closes the writable channel at the cost predicted by Corollary 1, and resistant judges keep near-zero lift; among measured offline committees, none Pareto-dominates its strongest single constituent.

2 Problem Formulation

2.1 Setup and Mixed Audit Views

Let an execution trace be $\tau = (E, U)$, where E represents trusted execution evidence (actions, arguments, tool receipts, environment states, timestamps, dependency edges) and U denotes agent-authored audit content (messages, explanations, first-person rationales). A task failure is indicated by $y(\tau) = \text{FAILURE}$. An automated attributor A maps a rendered audit view $\mathcal{V}(\tau)$ to a predicted responsible agent $\hat{a} \in \mathcal{A}$. Causal failure responsibility need not be unique; we denote the ground-truth reference set as $Z^\star \subseteq \mathcal{A}$. An innocent target $z_t \in \mathcal{A}$ satisfies $z_t \notin Z^\star$.

2.2 Targeted Blame Shifting

Definition 1 (Field-ownership write boundary)**.** For an agent $a \in \mathcal{A}$, let $\mathcal{W}_a$ be the set of field indices in U produced by a. An audit modification δ satisfies field ownership if $\delta(i) = 0$ for all fields $i \notin \mathcal{W}_a$.

Definition 2 (Task-execution-invariant attack)**.** An attack $\delta \in \Delta(\mathcal{W}_a)$ targeting victim z_t is task-execution-invariant if it alters only a's owned audit content while preserving trusted execution evidence, outcome, and ground-truth responsibility:

$$E(\tau \oplus \delta) = E(\tau), \quad y(\tau \oplus \delta) = y(\tau), \quad Z^\star(\tau \oplus \delta) = Z^\star(\tau). \tag{1}$$

We assume throughout that outcome and responsibility are functions of the execution record, $y = y(E)$ and $Z^\star = Z^\star(E)$, as in benchmark annotations that label the decisive agent from the execution; then every edit confined to U_a is task-execution-invariant. We instantiate this constrained class with ownership-checked narrative edits. A sample is statically valid under condition $\mathcal{C}(\tau, \delta)$ only if speaker-ownership and immutable-field equality both hold, and $z_t \notin Z^\star$. Invalid samples are excluded from every lift, defense and cross-judge comparison (HC/AG/TE n=54/47/51); the only rows on other task sets are the descriptive per-judge mechanism table and the n=30 placement subset, which state their n.

2.3 Blame Integrity of Audit Interfaces

Write $U = (U_a)_{a \in \mathcal{A}}$ with $U_a \in \mathcal{U}_a$ the content of a's owned fields, $\emptyset \in \mathcal{U}_a$ the empty narrative, and call the pair (φ, J) of an interface and a judge a *mechanism*.

Definition 3 (Audit interface)**.** An interface φ maps (E, U) to the judge-visible view; $\varphi_a(E, U)$ is the content it shows for a's fields. (i) *Flat*: $\varphi_a = U_a$. (ii) *Style-filtered* φ_g: a surface predicate g that does not consult E sets $\varphi_a = U_a$ if $g(U_a)$ and $\emptyset$ otherwise. (iii) *Citation-verified* φ_{cv}: retains only the claims in U_a entailed by E. (iv) *Trusted-only* φ_E: $\varphi_a = \emptyset$ for every a.

Definition 4 (Blame integrity)**.** Fix τ and $a \in Z^\star(\tau)$. The blame indicator of a is $\beta_\varphi(\tau, u) = \mathbf{1}[J(\varphi(E, U_{-a}, u)) = a]$ for $u \in \mathcal{U}_a$. The view-level reachable set is $\mathcal{R}_\varphi(\tau, a) = \{\varphi(E, U_{-a}, u) : u \in \mathcal{U}_a\}$. (φ, J) has *blame integrity* on a family $\mathcal{T}$ if $\beta_\varphi(\tau, u) = \beta_\varphi(\tau, U_a)$ for every $\tau \in \mathcal{T}$, $a \in Z^\star(\tau)$ and $u \in \mathcal{U}_a$; equivalently, the indicator $v \mapsto \mathbf{1}[J(v) = a]$ is constant on $\mathcal{R}_\varphi(\tau, a)$.

The indicator tracks whether a itself is blamed, so moving blame from a to another member of $Z^\star$ is also a violation. Because Definition 2 keeps E, y and $Z^\star$ fixed, a change in whether a is blamed under an ownership-legal edit is a failure of the mechanism rather than a response to a new trace (Lemma 1, Appendix A); a verdict shift between two agents other than a does not alter a's blame indicator.

Remark 1 (Strategic reading). Give the responsible agent a utility that depends on the verdict only through whether it is blamed and is non-increasing in β_φ, e.g. $u_a = 1 - \beta_\varphi$ (evasion). At fixed (E, U_{-a}), blame integrity makes every ownership-legal report yield the same own-blame utility, so reporting U_a is a weak best response (deviation-proof under blame avoidance). For targeted utility $\mathbf{1}[J(\cdot) = z_t]$, integrity rules out profitable targeting whenever the clean verdict names a, which is automatic for clean-correct verdicts on singleton-responsibility traces ($|Z^\star| = 1$). The converse fails, since integrity also forbids unprofitable changes, so we measure integrity and state strategyproofness only as its consequence (Appendix A).

Proposition 1 (Evidence-independent filters are bypassable or own-blame inert)**.** *Let φ_g be a style filter that rejects at least one message, and let $\mathcal{U}_a^g = \{u : g(u)\}$. Then on a's fields φ_g shows $U_a \in \mathcal{U}_a^g$ or $\emptyset$, and exactly one holds: (i)* bypassable*: for some τ, $a \in Z^\star(\tau)$, $\beta_{\varphi_g}(\tau, \cdot)$ is nonconstant on $\mathcal{U}_a$, so an ownership-legal edit alters whether a is blamed; (ii)* own-blame inert*: $\beta_{\varphi_g}(\tau, u) = \beta_{\varphi_g}(\tau, \emptyset)$ for all u, so on a's fields φ_g is blame-equivalent to deleting a's narrative.*

Corollary 1 (The robustness–utility tradeoff is forced)**.** *On singleton-responsibility traces ($|Z^\star| = 1$), if deleting the responsible agent's surviving narrative strictly lowers clean attribution on some trace, $\beta_{\varphi_g}(\tau, \emptyset) < \beta_{\varphi_g}(\tau, U_a)$, then φ_g is bypassable.*

Proposition 1 covers only admission rules that ignore E; it is not an impossibility result for evidence-dependent interfaces, which are classes (iii)–(iv). Case (i) guarantees that some admitted message changes the blame decision, not that a writer can find one aimed at a chosen z_t; that is an empirical question (Section 5.6).

Proposition 2 (Citation verification leaves selective disclosure)**.** *Identify every retained view with a canonical set of atomic claims, with non-claim phrasing, ordering, and framing removed. Let $C_a(E)$ be the E-entailed claims available to a, let φ_{cv} be sound and complete, and let every $S \subseteq C_a(E)$ be assertable by some $u \in \mathcal{U}_a$. Then the set of admitted claim sets is $2^{C_a(E)}$, and $(\varphi_{\mathrm{cv}}, J)$ has blame integrity if and only if the blame decision is invariant to the presented subset $S \subseteq C_a(E)$.*

Corollary 2 (Submission-independent disclosure)**.** *If the interface presents $C_a(E)$ irrespective of what a submits, then the retained view is a singleton and (φ, J) has a judge-independent blame integrity guarantee.*

Verification eliminates fabricated claims but preserves the choice of *which* true claims appear; a judge-independent guarantee results when the interface, not the agent, makes that selection (Milgrom, 1981; Grossman, 1981). Rather than claiming abstract game-theoretic complexity, these formal results provide an operational interface taxonomy that directly predicts the empirical findings: style filters remain bypassable (Section 5.6), eliminating narratives incurs a clean-accuracy cost where narrative is informative (Section 5.5), and verification removes non-entailed accusations but, applied to every turn, also costs clean accuracy (Appendix L; proofs in Appendix A).

2.4 Empirical Metrics

We report

$$\begin{aligned} \mathrm{TBSR} &= \Pr[A(\tau \oplus \delta) = z_t], & \mathrm{lift} &= \mathrm{TBSR}_{\mathrm{named}} - \mathrm{TBSR}_{\mathrm{noise}}, \\ \mathrm{AFR} &= \Pr[A(\tau \oplus \delta) \neq A(\tau)], & \mathrm{TRR} &= \Pr[A(\tau \oplus \delta) \in Z^\star], \end{aligned} \tag{2}$$

alongside clean agent accuracy ($\mathrm{CAA} = \Pr[A(\tau) \in Z^\star]$). Every evaluated trace has a single annotated responsible agent ($|Z^\star| = 1$ on all three splits), so the attacker a is that agent, $\mathrm{TRR} = \Pr[\hat{a} = a]$, and $1 - \mathrm{TRR}$ is the rate at which a escapes blame (Definition 4). The *matched noise* control appends one fixed, blame-neutral sentence ("the environment latency was high and screenshots were noisy"; no agent named) to the same attacker turn, and is scored against the same z_t. We also use the factorization $\mathrm{TBSR} = (1 - \mathrm{TRR}) \times \mathrm{steering}$, where steering is the probability of selecting z_t conditional on leaving $Z^\star$. TBSR measures *targeted* steering: it is neither identical to an integrity violation nor bounded above by it, but when the clean verdict is correct on singleton-responsibility

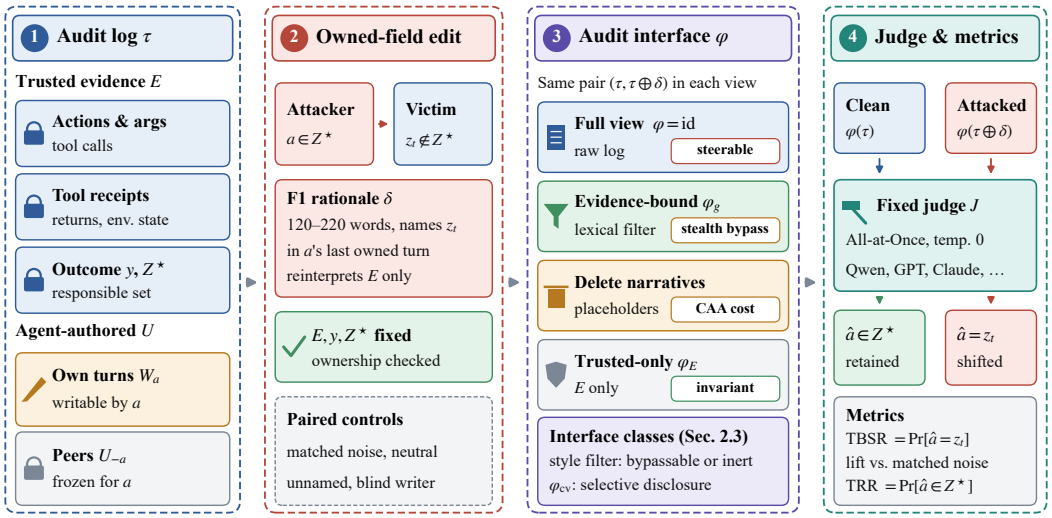

Figure 1: Threat model and paired evaluation views. Trusted evidence E is frozen; an ownership-legal named F1 edit rewrites only the attacker's writable fields W_a, producing $\tau \oplus \delta$ that names $z_t \notin Z^\star$. The same clean/attacked pair is scored under full, evidence-bound, delete-narratives, and trusted-only interfaces φ, and a fixed judge J attributes every view; paired controls share the task, victim and edited turn.

traces, steering blame away from a and onto z_t witnesses a violation. The violation rate itself is bounded below by $\Pr[\text{clean} = a \wedge \text{attacked} \neq a] \geq \text{CAA} - \text{TRR}$ on the same tasks (Remark 2). Cross-judge comparisons are restricted in advance to judges that are usable attributors on the clean task ($\text{CAA} \geq 0.40$) and to settings with $n \geq 50$; we call a transfer *successful* when the paired lift is at least $+0.20$ with a 95% CI excluding zero.

3 Threat Model

Figure 1 and Table 7 (Appendix G) fix the write boundary: trusted evidence E is frozen; the attacker may edit only its writable fields $W_a \subseteq U$. Main results apply an ownership-legal named F1 edit after execution with E and $Z^\star$ held fixed, then score the same clean/attacked pair under full, evidence-bound, delete-narratives, and trusted-only interfaces φ, plus a filter-aware stealth rewrite against evidence-bound (Sections 5.5 and 5.6).

The attacker knows the interface, sees trusted evidence, and receives $Z^\star$ and a preselected innocent victim z_t before writing (single-pass attacks do not query the judge). Victims are drawn from a verified innocent set; hindsight selection after querying the judge is excluded from the main tables. Access to $Z^\star$ is an oracle for measuring targeted steerability, not a claim about deployed agents; Section 5.2 removes it. The main attack writes the rationale into the attacker's existing final message turn; the same text appended as a new final message, which an append-only or signed log still admits, keeps most of the lift (Section 5.2).

4 Method

4.1 Attack Families

The attack leaves the execution trace unchanged and edits only post-execution text that the responsible agent is permitted to write. Every candidate edit must pass speaker-ownership and immutable-field equality checks before entering the TBSR evaluation set. We stratify edits by factuality. **F1** allows self-exoneration and selective interpretation of existing trusted events, but not fabricated receipts. **F2** allows a fabricated accusation and is reported only for comparison.

Split	Source	Input	Own.	No z_t	Valid n
HC	Who&When HC	58	3	1	54
AG	Who&When AG	50	0	3	47
TE	TraceElephant	58	7	0	51

Table 1: Evaluation subsets. "Own." counts tasks failing speaker-ownership or immutability checks. "No z_t" counts traces containing no innocent agent, for which no admissible target exists; both groups are removed from TBSR evaluation sets.

The main attack is named F1: (1) the responsible agent is the attacker a, and we sample $z_t \notin Z^\star$; (2) a short trusted-evidence excerpt is taken from the fixed execution record; (3) a language model writes a 120–220 word first-person rationale that names z_t and ends with a self-exonerating statement; (4) the rationale is appended only to a's last writable turn. Samples that violate ownership are excluded from the main metrics. F1 is a generation constraint (no invented tool returns, actions or observations; interpretive claims allowed), which writers follow imperfectly: a claim-level audit finds 3–9% of claims asserting events absent from the log (Appendix O). Prompts and the filter-aware variant are specified in Appendix C.

4.2 Defenses

Narrative deletion. Agent-authored text is replaced by fixed placeholders, closing the writable channel.

Evidence-bound attribution. A lexical style filter removes explicit audit blocks and blame-heavy lines while retaining operational text such as tool requests and status updates (Appendix D). Despite the name, it does not consult E: it is a class-(ii) interface in Definition 3, and the semantic, E-dependent alternative is the citation-verified view below.

Citation-verified / grounded probes. Appendix evaluations retain only E-supported content in every writable turn, or require the judge to cite trusted turns (Appendices L and P.6).

5 Experiments

5.1 Setup

Data. Table 1 lists the evaluation subsets: Who&When Hand-Crafted (HC) and Algorithm-Generated (AG) (Zhang et al., 2025), plus TraceElephant failed runs with static expert labels (Chen et al., 2026). Ownership-constrained edits keep execution fixed; matched noise, cross-writer, and volume controls accompany the attack.

Attributors. Unless noted, results use the Who&When All-at-Once interface with Qwen-Max and the published prompt (Zhang et al., 2025): a flat judge that reads the full mixed view in one pass. We also test Binary Search and Step-by-Step under the same base model, and two structured-prompt baselines (scope-then-localize; receipt cross-examination) that add localization or evidence-checking instructions under the same flat interface. Cross-judge transfer holds the attack text fixed and varies the attributor; further details appear in Appendices E and F.

Protocol. Every attack is paired with a matched noise control on the same tasks, victim, writer, placement, and judge. Defense comparisons reuse the same valid-victim task sets as Table 2 (n=54 / 47 / 51). Because primary targeted lifts ($+0.35$ to $+0.59$) are 3–5$\times$ their bootstrap interval half-widths, evaluated sample sizes ($n \approx 50$ per split) provide ample statistical power for confirmatory inference (family-wise adjusted $p \leq 0.012$). We report 95% paired bootstrap intervals over tasks (2000 resamples, fixed random seed); judges are queried through commercial model-serving interfaces with temperature set to zero.

Setting	CAA	Noise	Named (lift)	AFR	TRR
HC (n=54)	0.61	0.19	**0.78** ($+.59_{[.46,.72]}$)	0.61	0.09
AG (n=47)	0.68	0.15	**0.53** ($+.38_{[.23,.53]}$)	0.43	0.38
TE (n=51)	0.63	0.12	**0.53** ($+.41_{[.27,.55]}$)	0.41	0.37

Table 2: Main attack results under speaker-ownership filtering. Tasks with no admissible innocent victim are excluded. Intervals are 95% paired bootstrap CIs; all three lifts remain significant after Holm correction over the nine pre-specified primary tests enumerated in Appendix M (largest adjusted $p = 0.012$).

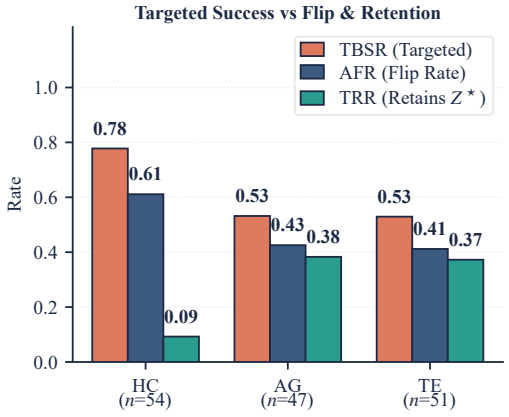

Figure 2: TBSR, AFR, and TRR under named F1. Targeted success rises while retention of $Z^\star$ falls.

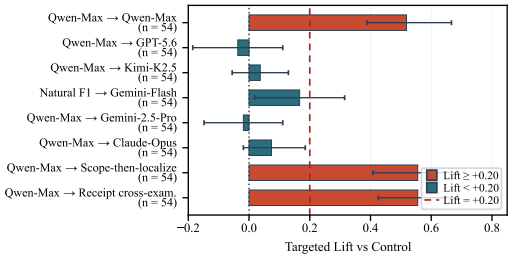

Figure 3: Cross-judge transfer and structured-prompt probes on the 54 valid-victim HC tasks (CAA ≥ 0.40; the Qwen-Plus writer row covers the 24 of them in its run). Bars show targeted lift over matched controls with 95% paired bootstrap intervals; the dashed line marks lift $+0.20$.

5.2 Main results

Table 2 and Figure 2 report the flat All-at-Once setting (Qwen-Max). Named F1 raises TBSR over matched noise on all three splits (HC $+0.59$, AG $+0.38$, TE $+0.41$), all with intervals excluding zero after Holm correction. Actions, tool receipts, and $Z^\star$ are held fixed; only attacker-owned audit text changes. AFR rises and TRR falls under the same edits. Measured directly against Definition 4, at least $\mathrm{CAA} - \mathrm{TRR} = 0.52$ / 0.30 / 0.26 of tasks turn from a correct clean verdict into an incorrect attacked one (Remark 2).

Displacement versus selection. TBSR conflates displacing a with selecting z_t, which coincide on traces with one innocent agent (51% of HC, 65% of TE). On traces with at least two innocent candidates the lift persists (HC $+0.46_{[.27,.65]}$, AG $+0.37_{[.24,.53]}$, TE $+0.28_{[.06,.50]}$), and the judge picks the precommitted victim above a per-trace uniform draw by $+0.27$ to $+0.30$: the edit both displaces blame and steers it.

Writer–judge transfer is in Figure 3. Under the same named F1 attack, GPT-5.6, Kimi-K2.5, Claude-Opus, and Gemini-2.5-Pro stay near matched noise (lifts -0.04 to $+0.07$; all CIs include zero). A natural-prose F1 variant yields a smaller Gemini-Flash effect (lift $+0.17$, 95% CI $[+0.02, 0.31]$). F2 exceeds noise but underperforms named F1, so the main effect does not need fabricated receipts. Qwen-Max scope-then-localize and receipt cross-examination keep clean CAA ($0.57/0.54$) yet lifts $+0.56/+0.56$ (Figure 3): under a fixed flat interface, prompt structure alone does not close the writable channel.

Writer–judge family overlap. Varying only the writer with the Qwen-Max judge, tasks, victims and control fixed rules out register matching (Table 13): Kimi-K2.5, Claude-Opus and GPT-5.6 reach

lifts $+0.65$, $+0.44$ and $+0.31$ (all CIs exclude zero) against $+0.52$ in-family; Kimi-K2.5 resists as a judge yet writes the strongest attack.

What carries the effect. One run scores five owned edits on the same 54 HC tasks with the same judge, each with the writer, word budget and two-position injection of named F1 (Table 16). A neutral recap that names no peer leaves TBSR and retention at noise level, so text volume does not drive the effect (Appendix N). A self-exonerating rationale that names no peer barely moves the verdict off $Z^\star$ ($+0.07_{[.00,.17]}$ over noise, TBSR lift $+0.06$ with CI covering zero), whereas naming a peer moves it by $+0.48_{[.35,.61]}$. Naming z_t is not enough either: a text that names it 1–4 times but clears it lowers its selection below noise ($-0.15_{[-.24,-.06]}$; Table 17), so the effect is the accusation, not name priming. Removing the oracle does not weaken the attack: a writer told neither that it is responsible nor whom to target, choosing its own scapegoat, lands the judge on that scapegoat at 0.87 (lift $+0.61_{[.48,.74]}$); on multi-innocent traces it picks our precommitted victim only 46% of the time, with lift $+0.62_{[.42,.81]}$. Nor does the shift depend on detected fabrications: on the subset of rationales that both automated auditors judge free of invented events, lifts remain $+0.62$ / $+0.48$ / $+0.35$ on HC / AG / TE, not detectably smaller than on all rationales (Table 18). Nor is rewriting historical turns strictly necessary: evaluated on HC with Qwen-Max, the same accusation appended as a new final self-report with no prior turn modified keeps lift $+0.44_{[.31,.57]}$.

Judge dependence. We decompose a targeted blame shift into two events on an attacked trace: the output leaves the true-responsibility set $Z^\star$ ($1 - \mathrm{TRR}$), and, conditional on leaving, selects the target. Across judges with CAA ≥ 0.40, TBSR and TRR exhibit a mechanical inverse trade-off ($\mathrm{TBSR}+\mathrm{TRR} \approx 0.85$, descriptive rank correlation $\rho = -0.99$ across the six unified judges; Figure 6 in Appendix I); the near-constant sum means the informative quantity is $P(z_t \mid \text{left } Z^\star)$. GPT, Kimi, and Claude retain $Z^\star$ more often ($\mathrm{TRR} = 0.61$–0.69 vs. 0.13 for Qwen-Max). Conditional target selection on multi-innocent traces where the judge leaves $Z^\star$ places Qwen-Plus and Gemini-Flash above chance, whereas GPT/Kimi/Claude stay near chance ($n_{\mathrm{cond}} = 13$–22). Table 4 shows the echo pattern: the judge reuses the attacker's accusation wording. The judge ordering is not specific to HC: on AG, named F1 lifts Qwen-Max by $+0.34$ but Kimi-K2.5 by $+0.08$, Gemini-2.5-Pro by $+0.04$ and Qwen3-Max by 0.00, and GPT-5.6 and Claude-Opus intervals cover zero (Table 10). On TraceElephant, against $+0.35$ for Qwen-Max on the same attack pack, Claude-Opus (-0.02), Gemini-2.5-Pro (0.00) and Kimi-K2.5 ($+0.08$) stay within or at the edge of noise, and GPT-5.6 shows a small but detectable lift ($+0.12_{[.04,.22]}$) below the $+0.20$ transfer threshold (Table 19). Near-zero lift means resistance to this edit, not correct attribution: resistant judges still miss $Z^\star$ on 31–40% of attacked HC traces.

5.3 Ablations and localization

Early-only edits stay near the matched control; end-of-trace named F1 carries the effect, and adding an early checkpoint adds no further lift (Table 20). F2 remains weaker than F1 under ownership checks (lift $+0.31$, 95% CI $[0.20, 0.44]$ vs. $+0.52$ for named F1 measured in the same ablation).

Single-pass named F1 stays near noise against Binary Search ($+0.11$, CI $[-0.02, +0.22]$) and Step-by-Step ($+0.02$, CI $[0.00, +0.06]$; Figure 4), whose leaf is often a turn the attacker cannot rewrite; an interactive adversary that conditions on intermediate binary-search decisions restores lift $+0.29$ (CI $[+0.15, +0.42]$).

5.4 Audit placement

We reuse the same F1 narrative and vary only its position on a matched HC subset (n=30; Table 22). An auxiliary trailing log outside the dialogue stays near noise (lift $+0.13$, CI covers zero), while the attacker's final attended dialogue turn recovers the effect (lift $+0.60$, 95% CI $[0.40, 0.80]$).

5.5 Defenses

Figure 5 compares matched views under the same named attack on HC, AG, and TraceElephant (n=54/47/51), using the same ownership exclusions as Table 2. Clean, matched-noise, and attacked conditions are scored under every view on the same tasks. On Who&When, evidence-bound attribution reduces attacked TBSR by 0.57 (HC) and 0.30 (AG) while clean CAA is unchanged; deleting all

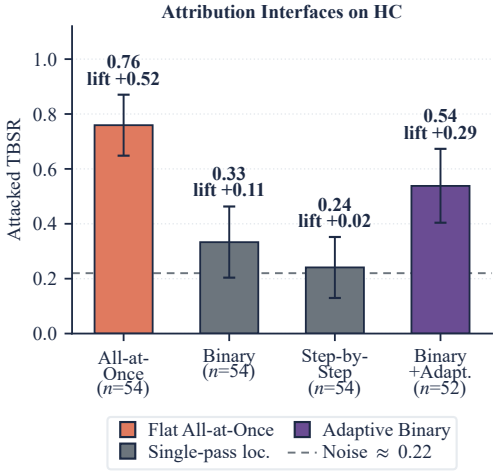

Figure 4: Attribution interfaces on HC (n=54; adaptive Binary n=52). Bars show attacked TBSR with 95% bootstrap intervals; the horizontal line is the matched single-pass noise rate. The All-at-Once bar (TBSR 0.76) comes from the localization run, a separate regeneration of Table 2's HC cell (0.78); replications differ by at most 0.074 in lift (Appendix M).

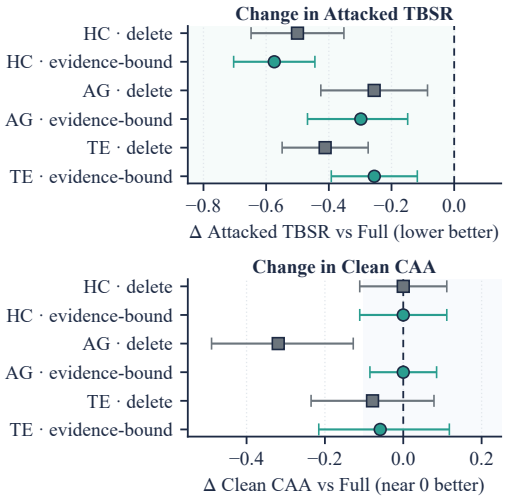

Figure 5: Paired defense effects vs. the full mixed view (HC n=54, AG n=47, TE n=51). Top: change in attacked TBSR (lower is better). Bottom: change in clean CAA (near zero is better).

narratives also reduces TBSR but costs 0.32 CAA on AG. On TraceElephant, both defenses reduce TBSR (delete Δ=−0.41, evidence-bound Δ=−0.25). The aggregate pattern is consistent with Corollary 1: deletion is free on HC, where narrative removal leaves clean accuracy untouched, and expensive on AG, where the judge uses narrative for clean attribution. Deleting only the responsible agent's narrative on AG costs 0.49 CAA (95% CI $[0.34, 0.62]$), versus 0.32 when all narratives are deleted. Exact per-view rows (Table 23) and a prompt-level grounding defense (Appendix P.6) appear in the appendix.

5.6 Filter-aware adaptive attack

Evidence-bound filtering is a style filter: it strips blame-heavy audit blocks while keeping tool requests and status updates. Proposition 1 predicts that such filters leave a cheap-talk channel, so a filter-aware attacker can rewrite blame as an operational status note under the same ownership boundary. Table 3 reports this stealth attack on all three splits under the same judge, exclusions, and matched-noise baselines as Table 2. On the full view, stealth is much weaker than named F1 (HC TBSR 0.24 vs. 0.78). Under the filter the HC ordering reverses: named F1 falls to lift +0.04 (CI covers zero) while stealth retains +0.17 [.07, .28], and filtering does not cut stealth relative to the full view (ΔTBSR +0.09 [−.02, .20]). TraceElephant shows the same pattern (named CI covers zero; stealth +0.08 [.02, .16]); on AG the filter leaves named F1 at +0.11 [.02, .19]. Every split therefore retains a positive ownership-legal lift under the filter, which is case (i) of Proposition 1; case (ii) appears as narrative deletion on HC, where clean accuracy is unchanged and named and noise inputs coincide. Against a length-matched control, stealth detection is near chance on blame cues and self-exoneration (AUC 0.52 and 0.50), whereas named F1 separates cleanly (Appendix N). Surface-form filtering and detection leave the ownership-legal channel open.

		Full		Evidence-bound	
Split	Attack	TBSR	Lift [CI]	TBSR	Lift [CI]
HC	named	0.78	$+0.59$ $[.46, .72]$	0.20	$+0.04$ $[-.06, .15]$
HC	stealth	0.24	$+0.06$ $[.00, .13]$	0.33	$+0.17$ $[.07, .28]$
AG	named	0.53	$+0.38$ $[.23, .53]$	0.23	$+0.11$ $[.02, .19]$
AG	stealth	0.32	$+0.17$ $[.06, .28]$	0.17	$+0.04$ $[-.06, .15]$
TE	named	0.53	$+0.41$ $[.27, .55]$	0.28	$+0.10$ $[-.02, .20]$
TE	stealth	0.18	$+0.06$ $[.00, .14]$	0.26	$+0.08$ $[.02, .16]$

Table 3: Filter-aware stealth attack on all three splits (Qwen-Max), on the same tasks and exclusion rule as Table 2 (n=$54/47/51$). Lifts are vs. matched noise on the same view; intervals are 95% paired bootstrap CIs. Clean CAA (full/evidence-bound) is $0.61/0.61$ (HC), $0.68/0.68$ (AG), $0.63/0.57$ (TE). Intervals covering zero indicate undetectable lift at this sample size; exploratory cells are uncorrected.

6 Related Work

Failure attribution. Who&When introduced agent- and step-level attribution for multi-agent failures (Zhang et al., 2025). Later clean-diagnosis work spans two-stage search (Sun et al., 2026), learned tracers (Zhang et al., 2026), causal graphs (Wang et al., 2026), fuller observability (Chen et al., 2026), hypothesis verification (Qiao et al., 2026), and intervention-style checks (Ma et al., 2026; Shah, 2026). These methods evaluate attribution on fixed logs; we study changes to agent-authored fields while execution records remain fixed.

Attacking judges and monitors. Chain-of-thought edits with actions held fixed flip success verdicts (Khalifa et al., 2026); monitors can be evaded via proxy agents (Isbarov & Kantarcioglu, 2026; Kutasov et al., 2025). LLM judges are steered by suffixes (Shi et al., 2024), universal phrases (Raina et al., 2024), candidate order (Wang et al., 2024), and self-preference (Panickssery et al., 2024). These attacks flip outcomes, evade detection, or shift candidate rankings. We instead fix the execution failure and evaluate targeted reassignment under an ownership boundary (Lemma 1; Table 5). Unnamed exoneration fails, and a Kimi writer is strongest against Qwen, so the effect is neither monitor evasion nor self-preference. Tamper-evident logs (Schneier & Kelsey, 1999) protect the record, not the truth of agent-authored narratives (Crawford & Sobel, 1982).

7 Discussion

The HC cost–robustness frontier (Appendix P.5) is dominated by the judge: Claude-Opus flat (CAA 0.65, lift $+0.07$) matches the best committee at a third of its cost and dominates Qwen-Max (0.61, $+0.59$). Style filters remain bypassable, so the operational choice is a resistant judge or a closed writable channel. Deletion is free where narrative is unused (HC) and costly where it is informative (AG -0.32), as Corollary 1 predicts.

Scope and limitations. Vulnerability is judge-dependent: flat Qwen judges are steerable, whereas other evaluated families resist this edit. Collusive multi-agent narratives, benign inconsistencies, and self-reports emitted during execution are out of scope. Learned attributors (SDBL, AgenTracer) lack public weights; public CHIEF shows no detectable lift (Appendix M).

8 Conclusion

Accusing an innocent peer in an agent's own narrative redirects flat judges on Who&When and TraceElephant without oracle knowledge or rewritten history—and persists on the subset not flagged by automated fabrication auditors—while other judge families resist. Style filters leave the channel open; deletion pays the cost of Corollary 1. Blame integrity is a property of the interface–judge pair and belongs in the specification of accountability mechanisms rather than post-hoc evaluation.

Ethics Statement

BlameShift studies manipulation of accountability infrastructure. All experiments run on public benchmark logs (Who&When, TraceElephant); no deployed attribution service is targeted. The released artifacts are the evaluation harness and records needed to reproduce the reported numbers. Every attack is reported together with the interfaces that block it. The intended use is the design of audit interfaces that remain invariant under ownership-legal narrative edits; we do not release a standalone attack tool.

Reproducibility Statement

Complete statements and proofs of all formal results are in Appendix A. Attack prompts, filter rules, judge interfaces, validation criteria, decoding settings, and seeds are specified in Appendices C–H. The supplementary code and data package contains the evaluation harness and the parsed per-record attribution decision behind every reported number, so every table and figure can be recomputed on CPU without API access.

AI Use Statement

In this work, we used generative AI tools to generate attack rationales (synthetic audit text), to implement the attribution judges and the entailment checker, and to edit the manuscript for grammar and phrasing. We have not used generative AI tools to develop the theoretical model or conceptual framework, to formulate or prove mathematical claims, to propose hypotheses, to design the experiments, or to interpret results. Attack and judge outputs are experimental measurements, scored against fixed labels; manuscript edits were checked by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Mengzhuo Chen, Junjie Wang, Fangwen Mu, Yawen Wang, Zhe Liu, Huanxiang Feng, and Qing Wang. Seeing the whole elephant: A benchmark for failure attribution in LLM-based multi-agent systems. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19888–19905, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.912. URL https://aclanthology.org/2026.acl-long.912/.

Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6): 1431–1451, 1982. doi: 10.2307/1913390.

Sanford J. Grossman. The informational role of warranties and private disclosure about product quality. *The Journal of Law and Economics*, 24(3):461–483, 1981. doi: 10.1086/466995.

Jafar Isbarov and Murat Kantarcioglu. Bypassing AI control protocols via agent-as-a-proxy attacks. *arXiv preprint arXiv:2602.05066*, 2026. URL https://arxiv.org/abs/2602.05066.

Navin Kartik, Marco Ottaviani, and Francesco Squintani. Credulity, lies, and costly talk. *Journal of Economic Theory*, 134(1):93–116, 2007. doi: 10.1016/j.jet.2006.04.003.

Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Sungryull Sohn, Yunxiang Zhang, Moontae Lee, Hao Peng, Lu Wang, and Honglak Lee. Gaming the judge: Unfaithful chain-of-thought can undermine agent evaluation. *arXiv preprint arXiv:2601.14691*, 2026. URL https://arxiv.org/abs/2601.14691.

Jonathan Kutasov, Yuqi Sun, Paul Colognese, Teun van der Weij, Linda Petrini, Chen Bo Calvin Zhang, John Hughes, Xiang Deng, Henry Sleight, Tyler Tracy, Buck Shlegeris, and Joe Benton. SHADE-Arena: Evaluating sabotage and monitoring in LLM agents. *arXiv preprint arXiv:2506.15740*, 2025. URL https://arxiv.org/abs/2506.15740.

Ming Ma, Jue Zhang, Fangkai Yang, Yu Kang, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. DoVer: Intervention-driven auto debugging for LLM multi-agent systems. In *The Fourteenth*

International Conference on Learning Representations, 2026. URL https://openreview.net/forum?id=mrEK16Jy6h.

Paul R. Milgrom. Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, 12(2):380–391, 1981. doi: 10.2307/3003562.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802, 2024. doi: 10.52202/079017-2197. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html.

Hezhe Qiao, Hanghang Tong, Ee-Peng Lim, Bing Liu, and Guansong Pang. VerifyMAS: Hypothesis verification for failure attribution in LLM multi-agent systems. *arXiv preprint arXiv:2605.17467*, 2026. URL https://arxiv.org/abs/2605.17467.

Vyas Raina, Adian Liusie, and Mark Gales. Is LLM-as-a-judge robust? Investigating universal adversarial attacks on zero-shot LLM assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7499–7517. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.emnlp-main.426/.

Bruce Schneier and John Kelsey. Secure audit logs to support computer forensics. *ACM Transactions on Information and System Security*, 2(2):159–176, 1999. doi: 10.1145/317087.317089.

Jaineet Shah. Causal agent replay: Counterfactual attribution for LLM-agent failures. *arXiv preprint arXiv:2606.08275*, 2026. URL https://arxiv.org/abs/2606.08275.

Jiawen Shi, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. Optimization-based prompt injection attack to LLM-as-a-judge. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, pp. 660–674, 2024. doi: 10.1145/3658644.3690291.

Kai Sun, Wenqiang Li, Bo Dong, Yuxin Lin, Jingyao Zhang, and Bin Shi. Scope delineation before localization: A two-stage framework for enhancing failure attribution in multi-agent systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40(39), pp. 33108–33116, 2026. doi: 10.1609/aaai.v40i39.40594. URL https://doi.org/10.1609/aaai.v40i39.40594.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, 2024. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Yawen Wang, Wenjie Wu, Junjie Wang, and Qing Wang. From flat logs to causal graphs: Hierarchical failure attribution for LLM-based multi-agent systems. *arXiv preprint arXiv:2602.23701*, 2026. URL https://arxiv.org/abs/2602.23701.

Guibin Zhang, Junhao Wang, Junjie Chen, Wangchunshu Zhou, Kun Wang, and Shuicheng Yan. AgenTracer: Who is inducing failure in the LLM agentic systems? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=l05DseqvuD.

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

A Formal Results and Proofs

This section proves the results stated in Section 2.3 and records the supporting lemma and remarks.

Setup. A trace is $\tau = (E, U)$ with trusted execution evidence E and agent-authored content $U = (U_a)_{a \in \mathcal{A}}$, where $U_a \in \mathcal{U}_a$ is the content of the fields owned by a (Definition 1) and $\emptyset \in \mathcal{U}_a$ (empty narrative). Ground-truth responsibility is a nonempty $Z^\star(\tau) \subseteq \mathcal{A}$. Interfaces act on owned fields separately (Definition 3); the lexical evidence-bound filter of Appendix D is of class (ii). We take J deterministic for readability; Remark 3 extends every claim to randomized judges.

Because $\beta_\varphi(\tau, u)$ depends on u only through $\varphi_a(E, U_{-a}, u)$, blame integrity (Definition 4) is equivalent to constancy of $\beta_\varphi(\tau, \cdot)$ on $R_\varphi(\tau, a)$. We therefore also write $\beta_\varphi(\tau, r)$ for $r \in R_\varphi(\tau, a)$ to mean the blame indicator under the view that places r in a's fields. Every argument below instantiates this observation for a different reachable set. Boundary cases with $R_\varphi(\tau, a)$ a singleton (in particular $\{\emptyset\}$) satisfy blame integrity by definition.

Lemma 1 (Well-posedness of the measurement)**.** *Let δ obey the field-ownership boundary of Definition 1 for some writer $a \in \mathcal{A}$ and be task-execution-invariant in the sense of Definition 2, and write $\tau \oplus \delta = (E,\ U_{-a},\ U_a \oplus \delta)$. Then E, y, $Z^\star$ and U_{-a} agree on τ and $\tau \oplus \delta$, so the two blame outcomes are evaluations of one function $\beta_\varphi(\tau, \cdot)$ at two points of $\mathcal{U}_a$. Hence $\beta_\varphi(\tau, U_a \oplus \delta) \neq \beta_\varphi(\tau, U_a)$ implies that (φ, J) violates Definition 4 at τ whenever $a \in Z^\star(\tau)$; conversely, if y and $Z^\star$ depend on τ only through E, every such violation is witnessed by an ownership-legal, task-execution-invariant edit of U_a.*

Proof. Definition 1 leaves U_{-a} untouched and Definition 2 leaves E, y and $Z^\star$ untouched, so the pair (E, U_{-a}) and the reference set $Z^\star(\tau)$ that define $\beta_\varphi(\tau, \cdot)$ are invariant under δ; only the evaluation point $u \in \mathcal{U}_a$ changes. A change of outcome is therefore non-constancy of $\beta_\varphi(\tau, \cdot)$ on $\mathcal{U}_a$, which for $a \in Z^\star(\tau)$ is exactly the negation of Definition 4. Conversely, a violation supplies some $u \in \mathcal{U}_a$ with $\beta_\varphi(\tau, u) \neq \beta_\varphi(\tau, U_a)$; the edit that replaces U_a by u and leaves every other field unchanged satisfies Definition 1 by construction; it leaves E unchanged, and because y and $Z^\star$ are functions of E it leaves them unchanged as well, so it satisfies Definition 2 (including the boundary edit $u = \emptyset$). Without that assumption an edit to U_a could in principle change the annotated outcome or responsibility, and the converse need not hold. □

By Lemma 1, a prediction flip under an ownership-legal edit is a property of the mechanism, not a change of the scored label.

Strategic reading (Remark 1). Let u_a depend on the verdict only through β_φ and be non-increasing in it. Under blame integrity, $\beta_\varphi(\tau, u) = \beta_\varphi(\tau, U_a)$ for all $u \in \mathcal{U}_a$, so u_a takes the same value at every ownership-legal report and no deviation is strictly profitable. Because β_φ indicates whether a itself is blamed, this includes deviations that pass blame to another responsible agent $b \in Z^\star$, $b \neq a$: such a deviation changes β_φ from 1 to 0 and is ruled out. For the targeted utility $\mathbf{1}[J(\cdot) = z_t]$, a deviation that moves a clean verdict naming a onto z_t likewise changes β_φ from 1 to 0; integrity therefore rules out profitable targeted deviations on traces whose clean verdict blames a, but not on traces where the clean verdict already names some other agent. The converse fails: a mechanism in which some edit moves the verdict *onto* a is strategyproof for an evasion-seeking agent but violates integrity. This is why we measure integrity, which does not depend on a utility model, and state strategyproofness only as a consequence.

Remark 2 (TBSR versus integrity violations). We state this remark for $|Z^\star| = 1$, which holds on every evaluated trace, so that $\hat{a} \in Z^\star$ if and only if $\hat{a} = a$. Targeted success requires both that the judge leave a and that it land on a preselected victim $z_t \notin Z^\star$, so $\mathrm{TBSR} \leq 1 - \mathrm{TRR}$. TBSR is not itself a violation rate: if the clean verdict already misses a, moving it from one innocent agent to z_t raises TBSR without changing β_φ. Violations of Definition 4 on a task set are instead witnessed by clean-correct, attacked-incorrect pairs, whose rate satisfies $\Pr[\text{clean} = a \wedge \text{attacked} \neq a] \geq \mathrm{CAA} - \mathrm{TRR}$. We report TBSR because it measures the stronger, attacker-directed event, and report CAA and TRR alongside it so that the violation bound can be read off. (If no innocent agent exists, no z_t is admissible and TBSR is undefined; such traces are excluded from the empirical evaluation.)

Remark 3 (Stochastic judges). For a randomized judge J, blame integrity requires equality of blame probabilities: $\bar{\beta}_\varphi(\tau, u) = \mathbb{E}[\beta_\varphi(\tau, u)] \in [0, 1]$ must be constant as a function of $u \in \mathcal{U}_a$. A single

realized output flip under a randomized judge does not by itself witness unequal blame probabilities. Evaluations using commercial model-serving endpoints with temperature zero approximate deterministic serving, and strategic reading applies in expected utility.

Proposition 3 (Flat interfaces)**.** *$(\varphi_{\mathrm{flat}}, J)$ has blame integrity on $\mathcal{T}$ if and only if $u \mapsto \beta_{\varphi_{\mathrm{flat}}}(\tau, u)$ is constant on $\mathcal{U}_a$ for every $\tau \in \mathcal{T}$ and every $a \in Z^\star(\tau)$.*

Proof. Under $\varphi_{\mathrm{flat}} = \mathrm{id}$ one has $R_{\varphi_{\mathrm{flat}}}(\tau, a) = \mathcal{U}_a$, so Definition 4 is precisely constancy of $\beta_{\varphi_{\mathrm{flat}}}(\tau, \cdot)$ on $\mathcal{U}_a$. □

Flat agent narratives are cheap talk in the sense of Crawford & Sobel (1982). Proposition 3 records the dichotomy for a fixed audit judge: the decision is constant over reachable messages, or some message manipulates it. How far a given judge can be moved is measured empirically in Appendix J, where two judges with statistically indistinguishable narrative reliance differ by $4.25\times$ in achievable lift.

Proof of Proposition 1. By Definition 3(ii), φ_a returns u when $g(u)$ holds and $\emptyset$ otherwise, so $R_{\varphi_g}(\tau, a) = \mathcal{U}_a^g \cup \{\emptyset\}$. Because g rejects some message, $\emptyset$ is attained as an image. Constancy of $\beta_{\varphi_g}(\tau, \cdot)$ on this image either holds for every τ and every $a \in Z^\star(\tau)$, or fails for some pair; these are exhaustive and mutually exclusive.

Case (i). Non-constancy supplies $r \in R_{\varphi_g}(\tau, a)$ with $\beta_{\varphi_g}(\tau, r) \neq \beta_{\varphi_g}(\tau, U_a)$ (equivalently, differing from the value at $\varphi_a(U_a)$). Any preimage $u \in \mathcal{U}_a$ of r is ownership-legal and changes the blame outcome, so Definition 4 fails.

Case (ii). Constancy on a set containing $\emptyset$ forces the common value to equal $\beta_{\varphi_g}(\tau, \emptyset)$, which is the outcome under φ_E on a's fields. If g rejects every message, $\mathcal{U}_a^g = \emptyset$ and $R_{\varphi_g}(\tau, a) = \{\emptyset\}$, a special case of (ii). □

The dichotomy is forced by g being a function of the message alone. Tightening or loosening g changes which messages survive, but cannot condition survival on support by E; escaping the two cases requires an interface of class (iii) or (iv) in Definition 3. Table 3 exhibits case (i) for the filter we deploy: a stealth rewrite survives g and still carries lift $+0.17$.

Proof of Corollary 1. Both $\emptyset$ and $\varphi_a(U_a)$ lie in $R_{\varphi_g}(\tau, a)$. Case (ii) of Proposition 1 would force $\beta_{\varphi_g}(\tau, \emptyset) = \beta_{\varphi_g}(\tau, U_a)$, contradicting the strict inequality, so case (i) holds. (If U_a already fails g, then $\varphi_a(U_a) = \emptyset$ and the antecedent is impossible; the hypothesis therefore applies only when the clean narrative survives the filter.) □

The hypothesis is per-trace, but a positive aggregate $\Delta\mathrm{CAA}$ under narrative deletion implies that at least one such τ exists, which is all the corollary needs. The tradeoff between attack suppression and clean accuracy is therefore predicted by the interface class, not an artifact of the particular lexical rules we implemented, and a filter that pays no utility cost is one whose retained narrative the judge was already ignoring. Our delete-narratives ablation removes every agent's narrative, whereas Corollary 1 constrains only the responsible agent's, so we measured that quantity directly with a third view that deletes narrative for $Z^\star$ alone. On AG (n=47, all three views scored together) clean attribution falls from 0.68 to 0.19, a drop of -0.49 (95% CI $[-0.62, -0.34]$), against -0.26 for deleting every narrative in the same evaluation (-0.32 in Table 23). The corollary's antecedent therefore holds on AG. The judge convicts partly on the responsible agent's own account, so silencing that agent alone leaves the innocent peers as the only narrative voices and is worse than silencing everyone. Appendix J reports the all-narrative values across splits, whose ordering matches the corollary: deletion is free on HC, where narrative removal leaves clean accuracy intact, and costly on AG, where it does not.

Assumptions of Proposition 2. Retained content is identified with its set of atomic claims. Soundness means φ_{cv} retains only E-entailed claims; completeness means it retains every E-entailed claim presented; and the message space is expressive: for every $S \subseteq C_a(E)$ some $u \in \mathcal{U}_a$ asserts exactly S (in particular $S = \emptyset$ via the empty narrative). In the boundary case $C_a(E) = \emptyset$ the reachable set is $\{\emptyset\}$ and the mechanism has blame integrity regardless of J.

Proof of Proposition 2. Soundness yields $\varphi_a(E, U_{-a}, u) \subseteq C_a(E)$ for every $u \in \mathcal{U}_a$. Completeness and expressivity give the converse: for each $S \subseteq C_a(E)$, submitting a message that asserts exactly S returns $\varphi_a = S$, since every claim of S is retained and nothing outside $C_a(E)$ survives. Hence $R_{\varphi_{\mathrm{cv}}}(\tau, a) = 2^{C_a(E)}$. Definition 4 is then exactly constancy of $\beta_{\varphi_{\mathrm{cv}}}(\tau, \cdot)$ on that power set. When $C_a(E) = \emptyset$, the power set is $\{\emptyset\}$ and constancy is automatic. □

Verification removes fabrication and leaves *selective disclosure of true statements* as the residual attack surface. Empirically, φ_{cv} removes the named-F1 lift, whose accusations E does not entail (Table 11); the residual selective-disclosure channel is shown formally here, and our stealth rewrite is too weak against the flat judge to probe it (Appendix L).

Proof of Corollary 2. A real-valued (or $\{0, 1\}$-valued) function on a singleton is constant, so Definition 4 holds. The empty-claim boundary $C_a(E) = \emptyset$ is included: the singleton is $\{\emptyset\}$. □

Corollary 2 is an audit analogue of full-disclosure results (Milgrom, 1981; Grossman, 1981). It obtains a judge-independent guarantee by having the interface determine which entailed claims appear, rather than only checking claims selected by the agent. A grounded-citation prompt (Appendix P.6) requires the judge to cite trusted evidence and falls back to a trusted-only view when it cannot, but the agent still chooses the emphasis, ordering and framing of the narrative it submits, so $R_{\varphi}(\tau, a)$ shrinks without collapsing to a singleton. Consistent with this residual choice, named TBSR falls from 0.76 to 0.54, but lift over matched noise remains $+0.48$ and clean CAA falls by 0.17 (Appendix P.6).

Credulity. A χ-credulous judge (Kartik et al., 2007) takes narrative blame claims at face value with probability χ and otherwise uses an E-measurable rule. The parameter χ indexes the two cases of Proposition 3. We fit it descriptively in Appendix I.

B Illustrative Examples and Closest-Work Contrast

Table 4 records two Who&When HC traces under the flat Qwen-Max All-at-Once judge. In the first, an owned end-of-trace F1 rationale that names Orchestrator flips the verdict from the true culprit (WebSurfer) and the judge echoes the accusation. In the second, naming Assistant misses and the judge lands on Orchestrator. Table 5 contrasts the threat with the closest prior attacks: those works flip an outcome, evade a monitor, or reorder candidates, whereas we hold the execution failure fixed and ask whether blame can be reassigned to a preselected innocent peer under a field-ownership write boundary.

Target selected (*Apple stock query*)	
Clean / attacked	WebSurfer / **Orchestrator** (z_t)
Attack excerpt	*"... Orchestrator made a critical mistake by not providing clear and specific instructions..."*
Judge rationale	*"The Orchestrator initially provided vague and insufficiently detailed instructions..."*
Target not selected (*APOD location query*)	
Intended vs. predicted	Assistant vs. Orchestrator

Table 4: Two HC examples for the Qwen-Max All-at-Once judge.

Work	Fixed failure?	Preselect victim?	Write boundary
Gaming Judge	No (outcome flip)	No	CoT only
Agent-as-Proxy	N/A (evasion)	No	Trace injection
BlameShift	Yes	Yes	Own audit fields

Table 5: Closest-work contrast with Khalifa et al. (2026) (Gaming Judge) and Isbarov & Kantarcioglu (2026) (Agent-as-Proxy).

C Attack Generation

All attacks are generated in a single pass under the field-ownership boundary; there is no iterative search over δ.

Named F1. The responsible agent a produces a first-person rationale of 120–220 words that names an innocent target z_t and argues that z_t made the decisive error. The text may reinterpret existing trusted evidence but may not invent tool returns, actions, or observations, and may not change the failure outcome. It ends with a fixed self-exoneration sentence and is appended only to a's last writable turn.

Filter-aware stealth F1. Under the same boundary, the writer produces an operational status note that still implicates z_t, but omits explicit blame vocabulary and attribution fields that the evidence-bound filter removes (responsibility cues, root-cause phrases, audit headers, and the self-exoneration sentence). Any remaining matched phrase is removed by a deterministic post-processing step.

D Evidence-Bound Filter

The evidence-bound view is a deterministic lexical filter. Trusted and execution-relevant turns are kept unchanged. Agent-writable turns retain operational content (for example, tool requests and status updates) and drop narrative content, including bracketed audit blocks, the self-exoneration sentence, and lines that match a fixed list of responsibility and blame cues. Turns that become empty after filtering are replaced by a fixed placeholder string. The delete-narratives view replaces all agent-writable content with a placeholder; the trusted-only view retains only trusted and execution-relevant turns.

E Attribution Interfaces

All-at-Once follows the published Who&When interface: the judge observes the task, the ground-truth answer, and the rendered audit view, then returns a responsible agent, a first mistaken step, and a short reason in a fixed three-field response format. Binary Search and Step-by-Step use the same model and the same prompt templates as Who&When; Step-by-Step inspects turns sequentially until the first positive error decision, and Binary Search recursively halves the trajectory. All judges are queried with temperature set to zero. Table 6 lists the attributors in the main comparison.

Attributor	Role
Who&When All-at-Once	Flat prompting baseline
Who&When Binary Search	Localization interface
Who&When Step-by-Step	Localization interface
Scope-then-localize	Structured prompt baseline
Receipt cross-examination	Structured prompt baseline
CHIEF (public release)	Hierarchical causal attribution

Table 6: Evaluated attributors in the main comparison.

F Structured-Prompt Baselines

We evaluate two Qwen-Max structured-prompt baselines to test whether additional prompt structure alone blocks the attack under a fixed flat interface.

Scope-then-localize. A two-stage procedure first nominates a short candidate set with a scope rationale and is instructed to ignore unsupported blame claims; the second stage then selects one agent and one first mistaken step from the same log and the output of the first stage.

Receipt cross-examination. A single-pass procedure first lists verified execution evidence (tool calls, arguments, and environment returns), then checks narrative blame claims against that evidence, and finally returns an agent and step decision based only on the verified evidence. Both baselines use the same full views, Qwen-Max model, temperature zero, tasks, and matched noise controls as the corresponding attacked conditions. On the 54 valid-victim tasks, clean CAA is $0.57/0.54$ and paired lifts are $+0.56\,[0.41, 0.69]$ and $+0.56\,[0.43, 0.69]$.

G Validation Protocol

Table 7 lists the field ownership used throughout.

Field	Trusted writer	Writable?
Own message / rationale	Agent	Yes
Other agents' messages	Other agents	No
Action / tool arguments	Orchestrator	No
Tool receipt / env. state	Tool / env.	No
Speaker, time, dependency	Log channel	No

Table 7: Field ownership corresponding to Figure 1. The final column indicates whether the attacker may rewrite the field.

Ownership and immutability. The ownership criterion permits only appending a new attacker-owned writable turn or editing an existing attacker-owned writable turn; any removed turn, changed speaker or role, or edit to a non-writable or non-owned turn fails. The immutability criterion requires that execution-relevant and non-writable turns are identical at the string level and that a checksum over immutable fields, together with the speaker sequence, is preserved. A sample is valid for static TBSR only if both criteria hold and $z_t \notin Z^\star$. Replay-based checks for execution faithfulness require an executable environment; reported Who&When results use the ownership and immutability criteria.

H Compute Resources and Randomization

Models and decoding. Judges and attack writers are queried with temperature zero and no sampling. Who&When and TraceElephant All-at-Once results use Qwen-Max; cross-judge transfer uses the judge named in each row of Table 9. The entailment checker for φ_{cv} is evaluated with Qwen-Max, with cross-family referee checks against Qwen-Plus. We release the parsed per-record attribution decision behind every reported number, so every metric is recomputable from the released records.

Randomization. Attack generation and bootstrap resampling (2000 draws over tasks) use fixed seeds. Every reported contrast pairs attacked, control, and clean conditions on the same task set.

Infrastructure. Inference uses remote model services; local analysis (scoring, bootstrap, plotting) is CPU-only and recomputes every table and figure from the released records in minutes.

I Judge-Heterogeneity Mechanism

Table 8 gives the per-judge decomposition summarized in Section 5. Cross-judge statistics use settings with CAA $\geq$ 0.40 ($\mathrm{TBSR} + \mathrm{TRR} \approx 0.85$, descriptive rank correlation $\rho = -0.99$ across the six unified judges, reflecting this mechanical trade-off rather than an inferential sample). DeepSeek and GLM have CAA below 0.40 and are excluded from these statistics. Table 9 collects all cross-judge, structured-prompt, ensemble and CHIEF rows.

J Narrative Reliance versus Manipulability

Corollary 1 predicts that a judge whose clean accuracy depends on agent-authored narrative cannot be protected by a style filter. We test the antecedent directly with a deletion probe on clean AG traces

Judge	n	CAA	TBSR	TRR	$P(z_t \mid \text{left})$	Cond. n
Qwen-Max	54	0.63	0.76	0.13	0.73 (0.45)	22
Qwen-Plus*	25	0.68	0.80	0.12	0.80 (0.48)	10
Gemini-Flash	54	0.54	0.43	0.46	0.62 (0.45)	16
Gemini-2.5-Pro	53	0.54	0.26	0.55	0.38 (0.46)	16
GPT-5.6	54	0.52	0.22	0.61	0.43 (0.44)	14
Kimi-K2.5	54	0.57	0.22	0.63	0.38 (0.45)	13
Claude-Opus	54	0.65	0.17	0.69	0.43 (0.43)	14
DeepSeek†	54	0.24	0.37	0.19	0.45 (0.46)	22
GLM-5.2†	54	0.15	0.11	0.11	0.16 (0.46)	25

Table 8: Per-judge mechanism on HC named F1 (descriptive). Except for the early Qwen-Plus probe ($*$, n=25), all rows use the 54 valid-victim tasks (Gemini-2.5-Pro has n=53 due to one gateway error). $P(z_t \mid \text{left } Z^{\star})$ reports targeted selection rate (parentheses: uniform chance among innocents) on multi-innocent traces where the judge left true responsibility (n_{cond} traces). †: CAA $<$ 0.40.

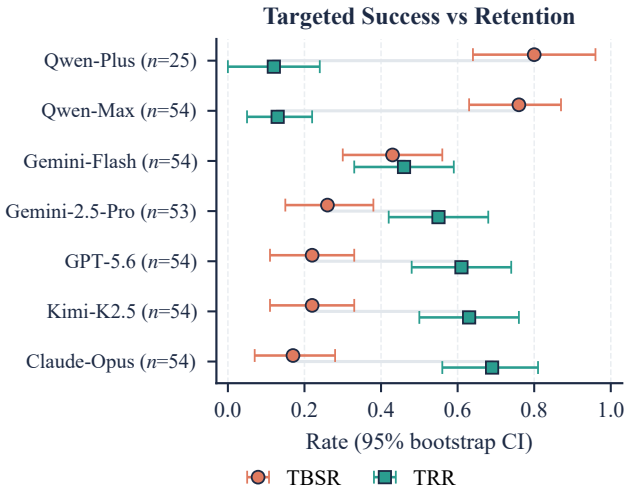

Figure 6: TBSR versus TRR across judges with CAA $\geq$ 0.40 under named F1 on HC. The two rates sum to 0.85 on average with descriptive rank correlation $\rho = -0.99$ across the six unified valid-basis judges. Points show 95% bootstrap intervals.

Writer	Judge	n	TBSR	Lift [CI]
Qwen-Max	Qwen-Max	54	0.76	$+0.52_{[.39,.67]}$
Qwen-Plus	Qwen-Max	24	0.79	$+0.54_{[.33,.75]}$
named	Scope-then-loc.‡	54	0.80	$+0.56_{[.41,.69]}$
named	Receipt cross-exam.‡	54	0.70	$+0.56_{[.43,.69]}$
named	GPT-5.6	54	0.22	$-0.04_{[-.19,.11]}$
named	Kimi-K2.5	54	0.22	$+0.04_{[-.06,.13]}$
named	Claude-Opus	54	0.17	$+0.07_{[-.02,.19]}$
named	Gemini-2.5-Pro	54	0.26	$-0.02_{[-.15,.11]}$
natural F1	Gemini-Flash	54	0.43	$+0.17_{[+.02,.31]}$
named	DeepSeek-Flash†	54	0.37	$+0.24_{[+.11,.37]}$
named	Qwen+GPT+Claude maj.	54	0.28	$+0.20_{[+.09,.33]}$
named	GPT+Claude+Kimi maj.	54	0.17	$+0.07_{[-.02,.17]}$
named	CHIEF (release)	54	0.35	$+0.06_{[-.06,.17]}$

Table 9: Cross-judge transfer, structured-prompt baselines, ensembles, and CHIEF results on the 54 valid-victim HC tasks (the Qwen-Plus writer run covers 24 of them). Intervals are 95% paired bootstrap CIs. †: clean CAA below 0.40; ‡: structured-prompt baselines.

(n=50, no attack): ΔCAA is the paired drop in clean agent accuracy when all agent-written narrative is replaced by placeholders. The second axis is the targeted lift of named F1 on the same split. The probe measures *reliance*; the attack measures *manipulability*.

Judge	ΔCAA (reliance)	Lift (manipulability)
Qwen-Max	$+0.30$ [.12, .48]	$+0.34$ [.20, .48]
Qwen-Plus	$+0.18$ [.00, .36]	$+0.24$ [.12, .36]
Kimi-K2.5	$+0.32$ [.14, .52]	$+0.08$ [.02, .16]
Gemini-2.5-Pro	$+0.18$ [.02, .34]	$+0.04$ [$-$.04, .12]
Qwen3-Max	$+0.06$ [$-$.12, .24]	$+0.00$ [$-$.08, .08]
Claude-Opus†	$+0.16$ [.00, .32]	—
GPT-5.6†	$+0.14$ [$-$.04, .32]	—

Table 10: Narrative reliance (ΔCAA under deletion, clean AG traces) and manipulability (named-F1 lift, AG). Both axes use the ownership-filtered AG set (n=50); the main table further excludes traces with no admissible victim (n=47). Intervals are 95% paired bootstrap CIs. Rows with both axes are ordered by lift. †: reliance axis only (lift interval covers zero on AG; HC lifts appear in Table 9).

The reliance point estimate is positive for all seven judges, and the interval excludes zero for three (Qwen-Max, Kimi-K2.5, Gemini-2.5-Pro): narrative deletion costs clean accuracy, matching the utility cost of aggressive filtering on AG ($\Delta\mathrm{CAA} = -0.32$ in Table 23). Gemini-2.5-Pro relies on narrative ($+0.18$ [.02, .34]) yet named F1 moves it by only $+0.04$ [$-$.04, .12], so reliance and manipulability come apart.

Reliance magnitude does not determine lift magnitude. Qwen-Plus and Gemini-2.5-Pro share a reliance point estimate ($+0.18$) yet their lifts differ by $6\times$ ($+0.24$ vs. $+0.04$); Qwen-Max and Kimi-K2.5 have reliance intervals that overlap almost completely ($+0.30$ vs. $+0.32$), yet their lifts differ by $4.25\times$ ($+0.34$ vs. $+0.08$). Whether a judge *uses* unverifiable narrative and whether an adversary can *steer* it are distinct properties. We therefore report judge-specific results and use the credulity parameter only as a descriptive summary.

K CHIEF Evaluation

We evaluate the publicly released CHIEF pipeline on the 54 valid-victim HC tasks under clean, named-attack and matched-noise conditions. Clean CAA is 0.54 (95% CI $[0.41, 0.67]$), named TBSR is 0.35, matched-noise TBSR is 0.30, and the paired named-vs-noise lift is $+0.06$ $[-0.06, +0.17]$, an interval covering zero. The probe uses the public release as shipped.

L Citation-Verified Interface Evaluation

The citation-verified view φ_{cv} checks each agent-writable turn against trusted execution evidence and keeps only supported content. A deployed interface does not know which turn was edited, so the primary evaluation applies the check to *every* writable turn of clean and attacked traces alike (HC, the 54 valid-victim tasks, Qwen-Max All-at-Once). The entailment checker uses Qwen-Max (with Qwen-Plus as cross-model referee, model–model disagreement rate $\epsilon \approx 0.033$ on n=120 calibration claims). To calibrate the checker against gold ground truth, two human annotators independently labeled all n=120 in-situ calibration claims against the trusted execution logs (inter-annotator agreement 96.7%, Cohen's $\kappa = 0.81$; binary agreement 99.2%, $\kappa = 0.88$), adjudicating disagreements by consensus. Because agent-authored turns in benchmark logs consist predominantly of deliberative rationales, directives, and conversational framing rather than verbatim execution recapitulation, grounded factual claims are naturally sparse in writable text (4 positive vs. 116 ungrounded/contradicted claims). Against this gold reference, the confusion matrix is TP = 4, FP = 2, FN = 0, TN = 114, yielding 98.3% overall binary accuracy (118/120, 95% Clopper–Pearson CI [94.1%, 99.8%]). All 4/4 grounded claims were retained in this in-situ set (recall = 100.0%, 95% CI [39.8%, 100.0%]; false-drop rate 0.0%), while 2/116 ungrounded narrative assertions were retained (false-retain rate 1.7%, 95% CI [0.2%, 6.1%]; precision 4/6 = 66.7%, 95% CI [22.3%, 95.7%]). Because the natural sparsity of grounded claims in writable turns limits the statistical power of in-situ recall, we

additionally evaluate a stratified, balanced held-out probe of n=50 claims (25 verified grounded execution statements from tool observation logs and 25 ungrounded attack rationales). On this balanced set, the checker achieves TP $= 24$, FP $= 1$, FN $= 1$, TN $= 24$, confirming high grounded recall ($24/25 = 96.0\%$, 95% CI $[79.6\%, 99.9\%]$), low false-retain rate ($1/25 = 4.0\%$, 95% CI $[0.1\%, 20.4\%]$), and 96.0% balanced accuracy (95% CI $[86.3\%, 99.5\%]$).

	Full view	Citation-verified	Δ (cv − full)
Clean CAA	0.65	0.57	$-0.07_{[-.24,.09]}$
Matched-noise TBSR	0.24	0.30	$+0.06_{[-.11,.22]}$
Named-F1 TBSR	0.78	0.31	$-0.46_{[-.61,-.31]}$
Stealth-F1 TBSR	0.22	0.28	$+0.06_{[-.09,.20]}$
Named-F1 lift	$+0.54_{[.41,.67]}$	$+0.02_{[-.04,.09]}$	
Stealth-F1 lift	$-0.02_{[-.09,.04]}$	$-0.02_{[-.11,.07]}$	

Table 11: Citation verification applied to every writable turn (HC, n=54). Full-view clean CAA (0.65) reflects an independent evaluation run on the 54 valid-victim tasks (baseline 0.61 in Table 2); paired comparisons (Δ) are computed strictly on the same run. Lifts are over matched noise in the same view; intervals are 95% paired bootstrap CIs.

Applied symmetrically, verification removes the named-F1 lift entirely at a clean-accuracy cost of 0.07 (interval covering zero), reflecting that stripping unverified narrative can remove context when that narrative aided clean attribution. The filter-aware stealth rewrite has no lift over noise against this flat judge even on the full view in this run, so these data neither confirm nor refute the selective-disclosure prediction of Proposition 2; that residual channel is established formally, and empirically only against the lexical filter (Table 3).

An earlier, *oracle-scoped* run checked only the known edited turns (Table 12, ownership set n=55). Because clean traces contain no edited turn, clean CAA is unchanged there by construction (0.62 under both views) and should not be read as evidence that verification is free.

Family	full	evidence-bound		citation-verified	
Named F1	0.75	0.20	$\Delta+0.55\,[0.42, 0.67]$	0.20	$\Delta+0.55\,[0.42, 0.67]$
Stealth F1	0.25	0.27	$\Delta-0.02\,[-0.13, 0.09]$	0.18	$\Delta+0.07\,[0.00, 0.16]$
Matched noise	0.18	0.22	$\Delta-0.04\,[-0.11, 0.02]$	0.20	$\Delta-0.02\,[-0.11, 0.07]$

Table 12: Oracle-scoped run: HC TBSR by view when only the edited turn is checked (ownership-filtered n=55). Δ is paired $\mathrm{TBSR}_{\mathrm{full}} - \mathrm{TBSR}_{\mathrm{view}}$ (positive values indicate lower TBSR than the full view). Bootstrap 95% CIs are in brackets.

In the oracle-scoped run, verification and lexical filtering suppress named F1 identically ($\Delta+0.55$ each) and leave stealth TBSR nearly unchanged; the same ordering appears on AG and TE (named $\Delta+0.26$ and $\Delta+0.25$, stealth $\Delta\,0.00$ and $\Delta-0.10$; ownership-only basis n=50/51). Since stealth has little lift over noise on the full view to begin with, the symmetric run above is the one we rely on.

M Threats to Validity and Scope of Statistical Claims

This section states the main scope restrictions and potential confounds.

Attacker knowledge. In the main tables the attacker is given $Z^\star$ and a preselected innocent victim z_t, and writes with trusted evidence in view. Single-pass attacks do not observe judge output. The blind variant of Appendix O removes both the responsibility oracle and the target, and is at least as effective.

Writer–judge family overlap. Because the largest lifts pair Qwen-written text with Qwen judges, register matching is a candidate explanation. We test it by holding the judge (Qwen-Max All-at-Once), the HC task set (n=54), the victim assignment and the matched noise control fixed, and varying only the writer. All three cross-family writers achieve large positive lifts with intervals strictly excluding

zero (Table 13), with Kimi-K2.5 exceeding the in-family writer ($+0.65$ vs. $+0.52$), ruling out register matching as the driver of the effect.

Writer	Noise	Named	Lift [95% CI]
Qwen-Max (in-family)	0.24	0.76	$+0.52$ [.39, .67]
Kimi-K2.5	0.22	**0.87**	$+\mathbf{0.65}$ [.52, .78]
Claude-Opus	0.26	0.70	$+0.44$ [.32, .57]
GPT-5.6	0.20	0.52	$+0.31$ [.15, .46]

Table 13: Writer transfer with the judge held fixed (Qwen-Max All-at-Once, HC, n=54). Only the model writing the F1 rationale varies; task set, victim assignment, placement and matched noise are identical across rows.

TBSR–TRR dependence. Because $\mathrm{TBSR} + \mathrm{TRR} \approx 0.85$ is nearly constant across judges with CAA ≥ 0.40, the descriptive rank correlation $\rho = -0.99$ is largely mechanical and should not be treated as independent inferential evidence for the credulity account. The informative quantity is the conditional selection probability $P(z_t \mid \text{left } Z^\star)$ in Table 8. The χ fit is likewise descriptive: shared-χ leave-one-out prediction is unstable when the mirror sums are nearly constant.

Attribution interfaces. The main attack targets All-at-Once flat prompting. Binary Search and Step-by-Step reduce named-F1 TBSR from 0.76 to 0.33 and 0.24; the adaptive Binary attack reaches 0.54 (n=52) under a stronger interactive adversary that conditions on intermediate judgments.

Multiple comparisons. The confirmatory family is pre-specified and contains nine tests: the named versus matched-noise comparison on each split, and the evidence-bound and delete-narratives views versus the full view under attack on each split. All nine are paired binary outcomes on the same tasks, so we use McNemar's exact test and apply Holm correction within the family. All nine remain significant at $\alpha = 0.05$; the largest adjusted p is 0.012 (AG, delete-narratives) and the three main attack tests have adjusted $p \leq 3.1 \times 10^{-4}$. Cross-judge and cross-writer rows use the reporting gate ($n \geq 50$, CAA ≥ 0.40, lift $\geq +0.20$, paired interval excluding zero); rows outside the confirmatory family are descriptive and are not corrected.

Replication. Regenerating attack text and reissuing judge queries on the same task sets changes targeted lift by at most 0.074 and attacked TBSR by at most 0.039. Reported intervals resample tasks.

Target multiplicity. On a trace with one innocent agent, hitting the precommitted victim and leaving $Z^\star$ are the same event, and such traces are 51% of HC, 18% of AG and 65% of TE. Pooled TBSR therefore mixes displacement with selection. The main text reports the restriction to traces with at least two innocent candidates, where the lift persists and conditional selection exceeds a per-trace uniform baseline. Traces with no innocent agent admit no valid z_t and are excluded (1 HC, 3 AG, 0 TE).

Text volume and placement. The named-F1 block is injected before and after the original turn content and adds a median of 1520 characters, whereas the matched noise control adds a median of 95 characters at one position. Matched noise therefore controls for victim, writer, turn and judge, but not for text volume or the number of injection points. Appendix N resolves this with a third condition matched on writer, word budget, audit framing and two-position injection that names no peer: it injects 13× more text than matched noise and changes TBSR by $+0.000$.

Scope. The evaluation targets automated attributors on public benchmarks. Metrics are agent-level; predicted steps are recorded for localization interfaces. Fabricated F2 attacks are weaker than owned-field F1 (0.52 vs. 0.76).

N Neutral Control and Detection Baselines

Volume-matched neutral control. Matched noise fixes the victim, writer, edited turn and judge, but it appends a single fixed sentence of 95 median characters, whereas named F1 inserts a rationale

before and after the original turn content for a median of 1520 injected characters. A length-matched neutral condition holds the framing fixed while naming no peer.

Design. The same writer (Qwen-Max), word budget (120–220 words), audit-rationale framing and two-position injection are held fixed; the rationale is instructed to recap only the attacker's own steps, to name no other agent, and to assign no responsibility. The neutral recap injects a median of 1216 characters against named F1's 1520. All three conditions are judged on the same tasks, so the contrasts are paired. *named−neutral* isolates the blame content and *neutral−noise* isolates text volume and placement.

Result. Table 14 reports the three conditions. Increasing the injected text from 95 to 1216 median characters, at the same two positions and in the same audit register, leaves TBSR unchanged ($0.185 \rightarrow 0.185$) and true responsibility retention (TRR) nearly unchanged ($0.667 \rightarrow 0.685$). The entire lift appears only when the rationale names a peer and assigns responsibility ($+0.574\,[0.444, 0.704]$ over the neutral control). Text volume, audit framing and injection multiplicity therefore contribute nothing measurable, and matched noise, despite being $13\times$ shorter, is an unbiased control for this judge. The collapse of responsibility retention under named F1 ($0.685 \rightarrow 0.130$) is the same event seen from the other side: the judge is not adding noise to its decision but relocating it onto the named victim.

Condition	Injected chars	TBSR	TRR
Matched noise	95	0.185	0.667
Neutral recap	1216	0.185	0.685
Named F1	1520	**0.759**	0.130
neutral − noise		$+0.000$ [.000, .000]	
named − neutral		$+\mathbf{0.574}$ [.444, .704]	

Table 14: Text volume versus blame content (HC, n=54, Qwen-Max All-at-Once). The neutral recap matches named F1 on writer, word budget, audit framing and two-position injection, and differs only in naming no peer and asserting no responsibility. TRR measures retention of true responsibility ($A(\tau \oplus \delta) \in Z^\star$).

Detection baselines. We ask whether an auditor could simply filter the injected text. Four zero-cost features are scored on the injected span: a blame-cue lexicon, peer naming, self-exoneration phrasing, and character length; the composite is their mean z-score. Table 15 reports AUC and TPR at FPR=0.10 against the neutral control, which is the informative negative class because it is matched on volume and register.

	named F1		stealth F1	
Feature	AUC	TPR@.10	AUC	TPR@.10
Blame cues	1.000	1.000	0.518	0.052
Self-exoneration	1.000	1.000	0.500	0.017
Peer naming	0.991	0.983	0.991	0.983
Length	0.833	0.431	0.000	0.000
Composite	1.000	1.000	0.522	0.103

Table 15: Detecting the injected span against a length- and register-matched neutral control (HC, n=58). Named F1 separates on content features; the filter-aware stealth variant is near chance except on peer naming.

Detection scope. Named F1 is lexically conspicuous: a blame-cue lexicon separates it perfectly even from a length-matched control, so a deployment that anticipates this fixed attack can filter it. Stealth F1 is at chance on blame cues (0.518), self-exoneration (0.500) and the composite (0.522), yet under the lexical filter it retains lift $+0.17\,[0.07, 0.28]$ over matched noise (Table 3). Length is not a usable feature: its direction reverses across variants (0.833 for named, 0.000 for stealth), so no fixed threshold catches both. The single feature that separates stealth is peer naming, and peer

naming cannot be filtered without removing the interface's function, since attribution is precisely the act of naming an agent. Surface predicates that catch named F1 miss the filter-aware rewrite; peer naming detects it only by removing the attribution interface's defining operation.

O NAMING, PLACEMENT, ATTACKER KNOWLEDGE, FABRICATION, AND TE TRANSFER

Design. Two attack families are added to the HC pack of Appendix N and scored in one run together with matched noise, the neutral recap and named F1 (Qwen-Max All-at-Once, temperature zero, the 54 valid-victim tasks). Both use the named-F1 writer (Qwen-Max), word budget, first-person audit framing, exoneration sentence and two-position injection. *Unnamed exoneration* argues that the attacker's own step was not the root cause and may appeal to task ambiguity, missing context or tool limits, but is forbidden to name or allude to any other agent; no generated text named a peer. *Blind* tells the writer only that the task failed, that an auditor will assign responsibility, that it wishes to avoid blame, and the names of the other agents; it is not told that it is responsible, which agents are innocent, or whom to target, and it outputs its chosen scapegoat before the rationale. TBSR for the blind family is scored against that self-chosen target, and its matched control is the noise condition scored against the same target. One of 58 blind outputs named no parseable agent; all 54 valid-victim tasks parsed, and every chosen target was innocent.

Edit	TBSR	AFR	TRR	TBSR lift vs. noise	Left-$Z^\star$ lift vs. noise
Matched noise	0.22	0.13	0.61	—	—
Neutral recap	0.20	0.13	0.65	$-0.02_{[-.06,.00]}$	$-0.04_{[-.09,.00]}$
Unnamed exoneration	0.28	0.19	0.54	$+0.06_{[-.02,.13]}$	$+0.07_{[.00,.17]}$
Named F1	0.76	0.57	0.13	$+0.54_{[.41,.67]}$	$+0.48_{[.35,.61]}$
Blind (own target)	**0.87**	0.59	**0.09**	$\mathbf{+0.61}_{[.48,.74]}$	$\mathbf{+0.52}_{[.39,.65]}$

Table 16: Five owned edits in one run (HC, n=54, clean CAA 0.61). "Left-$Z^\star$" is $1 - \text{TRR}$. Blind TBSR and its lift use the writer's self-chosen target; on the 26 multi-innocent traces that target equals the precommitted victim in 46% of cases and the lift is $+0.62_{[.42,.81]}$. Intervals are 95% paired bootstrap CIs.

Reading. Displacement and targeting come together: an exculpatory narrative that points nowhere leaves the judge near its noise-level verdict, while the same budget spent naming a peer moves both retention and targeted success. The blind writer is at least as effective as the oracle-informed one, so the oracle in the main tables makes the measurement targeted rather than making the attack possible.

Victim-name control and append-only placement. A second run on the same 54 tasks and judge adds two conditions (Table 17). The *victim-named* control uses the named-F1 writer, word budget and two-position injection, instructs the writer to mention z_t twice (generated texts contain 1–4 mentions), states that z_t acted correctly and did not cause the failure, and blames no agent. The *append-only* condition takes the identical named-F1 text and emits it once as a new attacker message after the final turn, modifying no existing turn; its matched noise is appended the same way. Unlike the auxiliary trailing log of Section 5.4, the appended message is an ordinary turn of the dialogue.

Edit	TBSR	TRR	TBSR lift vs. control	Control
Matched noise (in turn)	0.22	0.67	—	—
Victim-named, cleared	0.07	0.78	$-0.15_{[-.24,-.06]}$	in-turn noise
Named F1 (in turn)	0.76	0.13	$+0.54_{[.41,.67]}$	in-turn noise
Named F1 vs. victim-named	—	—	$+0.69_{[.56,.81]}$	victim-named
Matched noise (appended)	0.22	0.63	—	—
Named F1 (appended)	0.67	0.24	$+0.44_{[.31,.57]}$	appended noise

Table 17: Victim-name control and append-only placement (HC, n=54, Qwen-Max All-at-Once, one run; clean CAA 0.65). Appended versus in-turn named F1 differ by $-0.09_{[-.19,.00]}$ in TBSR. Intervals are 95% paired bootstrap CIs.

While this cleared condition introduces reverse semantic polarity and functions as a strong anti-priming control rather than an isolated lexical ablation, naming z_t while clearing it lowers its selection below noise ($-0.15_{[-.24,-.06]}$), confirming that mere exposure to or copying of the victim's name does not explain the redirection. The attack also does not require rewriting history: evaluated on HC with Qwen-Max, a final self-report appended to an append-only log retains most of the lift ($+0.44_{[.31,.57]}$), which a signed or append-only log would not prevent. We did not test self-reports emitted during execution, which would require re-running the multi-agent systems.

Fabrication audit. Each named-F1 rationale behind the primary-run tables is split into atomic claims by an auditor model that sees the full clean log, and every claim is labelled as a supported event, a fabricated event (an action, tool output, observation or message content that the log does not contain or contradicts), or an interpretive claim; counterfactual and normative statements count as interpretive. Two auditors from different families (Qwen-Plus and Kimi-K2.5, temperature zero) label independently. The fabrication-free subset keeps only rationales that neither auditor flags.

Split	n	Claims	Fab. claims	≥ 1 fab.	κ	Free n	Lift (all)	Lift (free)
HC	54	7.8	8.6% / 9.4%	54% / 41%	0.45	21	$+0.52_{[.39,.67]}$	$+0.62_{[.43,.81]}$
AG	47	8.3	7.9% / 7.8%	28% / 38%	0.19	23	$+0.36_{[.21,.51]}$	$+0.48_{[.30,.70]}$
TE	51	7.7	2.8% / 8.1%	20% / 41%	0.17	26	$+0.35_{[.24,.49]}$	$+0.35_{[.19,.54]}$

Table 18: Claim-level fabrication audit of named F1 (Qwen-Plus / Kimi-K2.5). "Claims" is the mean number of atomic claims per rationale; "≥ 1 fab." is the share of rationales with at least one fabricated event; κ is Cohen's kappa on that flag. Lifts are paired named-vs-noise lifts in the primary attack run, on all rationales and on the subset neither auditor flags.

Roughly two thirds of claims are interpretive and 3–9% assert an event the log does not show, so F1 as generated is mostly but not entirely interpretive. Auditor agreement on whether a rationale contains any fabrication is moderate on HC and low on AG and TE, which is why we condition on the intersection. On that conservative subset the lift is no smaller than on all rationales, demonstrating that the attribution shift persists on the subset of rationales not flagged by either automated auditor.

Cross-judge transfer on TraceElephant. Table 19 runs the four resistant HC judges on the TE attack pack used for the Qwen-Max replication run, on the same 51 valid-victim tasks (Gemini-2.5-Pro: 50, one task lost to repeated gateway errors).

Judge	n	CAA	Noise	Named	Lift [CI]	TRR
Qwen-Max	51	0.63	0.14	0.49	$+0.35_{[.24,.49]}$	0.43
Kimi-K2.5	51	0.61	0.20	0.27	$+0.08_{[.00,.18]}$	0.65
GPT-5.6	51	0.51	0.14	0.25	$+0.12_{[.04,.22]}$	0.63
Claude-Opus	51	0.57	0.24	0.22	$-0.02_{[-.12,.08]}$	0.71
Gemini-2.5-Pro	50	0.50	0.20	0.20	$+0.00_{[-.08,.08]}$	0.58

Table 19: Named F1 on TraceElephant with the attack text fixed and the judge varied. The Qwen-Max row is the primary TE attack run (Table 2 reports the defense-confirmation run of the same pack).

P Additional Empirical Results

This section reports additional paired intervals and graphical summaries. Qualitative examples and the closest-work contrast appear in Appendix B.

P.1 Attack-factor ablation

Setting	Control	TBSR	Lift [CI]
Early only	0.24	0.27	$+0.04_{[.00,.09]}$
End-of-trace (F1)	0.24	**0.76**	$+\mathbf{.52}_{[.39,.67]}$
Early and end	0.24	0.74	$+.50_{[.37,.63]}$
Fabricated (F2)	0.20	0.52	$+.31_{[.20,.44]}$

Table 20: Attack ablations on the 54 valid-victim HC tasks with All-at-Once (Qwen-Max). Intervals are paired bootstrap CIs; F2 uses a separate matched control.

P.2 Localization interfaces

Setting	TBSR	Lift [95% CI]
All-at-Once named	0.76	$+0.52\ [0.39, 0.67]$
Binary × named	0.33	$+0.11\ [-0.02, 0.22]$
Step-by-Step × named	0.24	$+0.02\ [0.00, 0.06]$
Binary × adaptive	0.54	$+0.29\ [0.15, 0.42]$

Table 21: Localization results on HC (Qwen-Max). Binary and Step-by-Step use n=54; Adaptive Binary uses n=52 after the same ownership exclusions.

P.3 Audit placement

Locus	TBSR	Noise	Lift [CI]
Final attended dialogue turn	0.73	0.13	$+0.60_{[.40,.80]}$
Auxiliary trailing log	0.27	0.13	$+0.13_{[-.07,.33]}$

Table 22: Matched placement comparison on HC (n=30), using the same F1 narrative. Intervals are 95% paired bootstrap CIs.

P.4 Defense utility–robustness tradeoff

Split	View	CAA	TBSR	ΔTBSR	ΔCAA
HC	full	0.61	0.78	—	—
HC	delete	0.61	0.28	$-0.50_{[-.65,-.35]}$	0.00
HC	evidence-bound	0.61	**0.20**	$-\mathbf{0.57}_{[-.70,-.44]}$	0.00
AG	full	0.68	0.53	—	—
AG	delete	0.36	0.28	$-0.26_{[-.43,-.09]}$	-0.32
AG	evidence-bound	0.68	**0.23**	$-\mathbf{0.30}_{[-.47,-.15]}$	0.00
TE	full	0.63	0.53	—	—
TE	delete	0.55	**0.12**	$-\mathbf{0.41}_{[-.55,-.27]}$	-0.08
TE	evidence-bound	0.57	0.27	$-0.25_{[-.39,-.12]}$	-0.06

Table 23: Matched defense views under named F1 (HC n=54, AG n=47, TE n=51; same valid-victim task basis as Table 2). Δ is relative to the full view on the same tasks (95% paired CI); "—" marks the reference row.

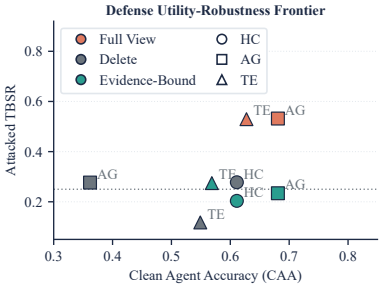

Figure 7: Utility–robustness frontier under the named-F1 attack. Points show clean CAA against attacked TBSR for the full, delete-narratives, and evidence-bound views on each split; lower and farther right is preferred.

P.5 Cost–robustness frontier

Table 24 compares every measured HC operating point by clean accuracy, targeted success, and judge queries per attribution. All rows use the admissible HC set (n=54); ensemble predictions are restricted to the same tasks before majority voting. The delete-narratives row has a lift interval of $[0, 0]$ because that view makes the named and matched-noise inputs identical, so the row satisfies blame integrity by construction—Proposition 1 case (ii)—and it is free here only because HC narrative was unused for clean attribution.

Operating point	Cost	CAA	Noise	Named	Lift [CI]
Qwen-Max, flat	1	0.61	0.18	0.78	$+0.59_{[.46,.72]}$
Qwen-Max, ev.-bound	1	0.61	0.17	0.20	$+0.04^{\star}_{[-.06,.15]}$
Qwen-Max, delete	1	0.61	0.28	0.28	$+0.00^{\star}_{[.00,.00]}$
Qwen-Max, gr. cite	2	0.46	0.06	0.54	$+0.48_{[.33,.63]}$
GPT-5.6, flat	1	0.52	0.26	0.22	$-0.04^{\star}_{[-.19,.11]}$
Claude-Opus, flat	1	**0.65**	0.09	**0.17**	$+0.07^{\star}_{[-.02,.19]}$
Kimi-K2.5, flat	1	0.57	0.18	0.22	$+0.04^{\star}_{[-.06,.13]}$
Cmte. G+C+K	3	0.65	0.09	0.17	$+0.07^{\star}_{[-.02,.17]}$
Cmte. Q+G+C	3	0.69	0.07	0.28	$+0.20_{[.09,.33]}$
Cmte. 5-way	5	0.67	0.11	0.26	$+0.15_{[.04,.26]}$

Table 24: HC cost–robustness frontier (n=54). Cost is judge queries per attribution. Committee rows are majority votes over GPT-5.6, Claude-Opus and Kimi-K2.5 (G+C+K) and over Qwen-Max, GPT-5.6 and Claude-Opus (Q+G+C). $\star$ marks a lift interval covering zero. Intervals are 95% paired bootstrap CIs.

Among robust operating points (lift interval covering zero), only single-query points are Pareto-efficient: Claude-Opus flat if narrative is retained, and Qwen-Max with narrative deletion if the judge is fixed. While the three-way committee Q+G+C reaches slightly higher clean CAA (0.69), it re-introduces vulnerability ($+0.20$ lift) at triple the cost. The robust committee G+C+K merely reproduces its best member (Claude-Opus) at three times the cost. The frontier is reported for HC; AG and TE defense tables corroborate the same interface tradeoff.

P.6 Grounded-citation defense probe

We test a grounded-citation approximation to the submission-independent interface in Corollary 2. The judge sees the mixed view with every turn tagged trusted or agent-writable and must cite Evidence Turns intersecting the trusted set, falling back to a trusted-only re-query otherwise. Because the agent still chooses what to write, $R_{\varphi}(\tau, a)$ shrinks without collapsing to a singleton (HC, n=54, Qwen-Max): named TBSR falls $0.76 \to 0.54$ ($\Delta - 0.22$ $[-.35, -.09]$) but matched noise falls with it, leaving lift $+0.48$ $[.33, .63]$ against $+0.52$ $[.39, .67]$ flat, at a CAA cost of $0.63 \to 0.46$ driven by a 0.33 fallback rate. Consistent with Corollary 2, absolute TBSR falls while targeted lift over matched noise remains large.