LOCALIZATION IS NOT ATTRIBUTION: DIAGNOSING WHO-AND-WHEN BLAME IN MULTI-AGENT LLM SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

When a multi-agent LLM system fails, two questions must be answered to enable repair: *when* did the trajectory go wrong, and *who* is responsible. We show these are far from equally hard. Across four frontier LLM judges on the WHO&WHEN benchmark of *real* multi-agent failures, every method we test—transcript-reading baselines, structured step-finding, and a strong leave-one-out evidence attribution—localizes the failing step far more often than it correctly attributes the responsible agent. We quantify this gap with a diagnostic metric, right-step-wrong-agent (rsw): the fraction of correctly localized cases in which the blamed agent is nonetheless wrong. rsw stays stubbornly high (0.19–0.24) for all methods and, crucially, *rises* on the annotation subsets that multiple judges agree are most trustworthy—evidence that it reflects a genuine model limitation, not label noise. We trace the error to *role-charter collapse*: when two agents play overlapping roles, judges summarize them into indistinguishable profiles and then misattribute blame. We propose **Role-Op Attribution (ROA)**, which builds *distinctive* charters via leave-one-out contrast and aligns them to per-agent operation signatures. ROA is the first method to significantly reduce rsw (-0.036; 95% CI $[-0.068, -0.003]$, paired bootstrap, n=632) while nudging joint who+when accuracy upward, at 40% lower judge-input cost than the strongest baseline. A component ablation and a self-consistency control show the fix is structural rather than a matter of more sampling, and the same localization–attribution dissociation generalizes to vision–language agents. We release all code and per-case verdicts.

1 INTRODUCTION

Multi-agent LLM systems now coordinate tool use, code execution, retrieval, and verification across specialized agents (Wu et al., 2024; Hong et al., 2024; Yao et al., 2023). When a run fails, engineers and repair loops need to identify both the responsible agent and the decisive step—the failure attribution problem (Zhang et al., 2025b; Cemri et al., 2025; Zhang et al., 2025a). Recent work studies this problem on WHO&WHEN, which pairs real failures with human annotations, but reported accuracies remain low and the cause is unclear (Zhang et al., 2025b).

We show that the two parts of attribution are unequally difficult. Across four frontier judges and five methods, we measure the rate at which a method gets the decisive step right but blames the wrong agent, which we call right-step-wrong-agent (rsw). Rsw remains high and flat (0.19–0.24) across methods: judges often identify *when* a run derailed but not *who* caused it. Gold verification rules out annotation noise: rsw rises from 0.210 overall to 0.465 on unanimously verified labels, indicating a genuine reasoning limitation.

Inspecting these failures, we identify *role-charter collapse*. Agents in real trajectories often perform overlapping operations, so a judge's summaries become near-identical and the judge defaults to a salient-but-wrong agent. Step localization, by contrast, reads the transcript directly and is less affected by this collapse.

Guided by this diagnosis, we propose **Role-Op Attribution (ROA)**. ROA (i) builds *distinctive* charters by contrasting each agent against the others in a leave-one-out fashion, forcing the description to capture what makes an agent unique, and (ii) grounds those charters in per-agent *op-*

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

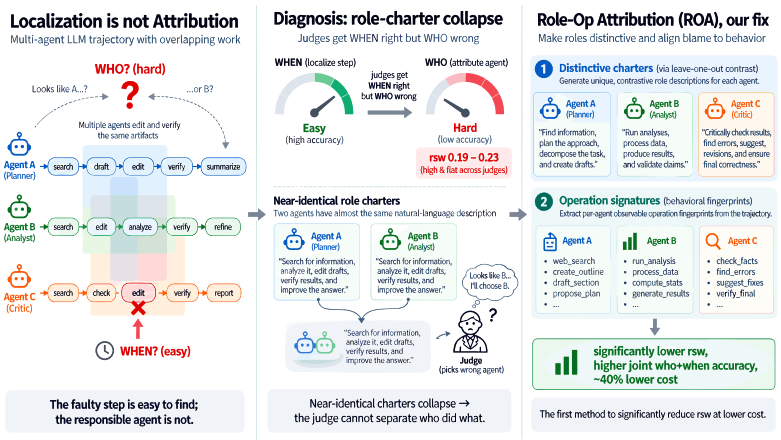


Figure 1: Overview of **ROA**. **Left:** in a multi-agent run the faulty *step* (*when*) is easy to localize, but attributing the responsible *agent* (*who*) is not. **Center:** judges score *when* high yet *who* low—the right-step-wrong-agent rate (rsw) stays flat at 0.19–0.23—because near-identical role charters *collapse*, so the judge picks a salient-but-wrong agent. **Right:** ROA restores separability via (i) distinctive charters (leave-one-out contrast) and (ii) per-agent operation signatures, the first method to significantly reduce rsw and raise joint accuracy at lower cost.

eration signatures extracted from the trajectory, so that blame is aligned to observable behavior rather than a paraphrase. ROA is the first method in our study to *significantly* reduce rsw (-0.036; 95% CI $[-0.068, -0.003]$, paired bootstrap, n=632) while nudging joint who+when accuracy upward ($+0.027$; 95% CI $[-0.003, 0.059]$), and it does so at 40% lower judge-input cost than the strongest baseline. We are careful here: the rsw reduction is the robust, significant effect (CI excludes zero); the joint gain is positive and consistent in direction but its CI edges zero, so we treat rsw as the primary evidence and joint as corroborating.

Two controls show ROA is not a generic "more compute" effect. A component ablation reveals the two mechanisms are *coupled*: adding distinctive charters alone *hurts* (rsw $0.210 \rightarrow 0.230$), and only their combination with operation alignment wins. And self-consistency voting over five diverse samples *fails* to reduce rsw (it slightly worsens to 0.215), because collapse is a systematic blind spot that repeated sampling only reinforces. Attribution needs a structural intervention, not more draws.

Contributions.

- **Phenomenon and root cause.** We introduce right-step-wrong-agent (rsw) and show, across four judges and five methods, that *localization is not attribution*: real multi-agent blame fails specifically at the "who" step. Via a gold-verification study we further establish that this gap is a genuine model limitation rather than label noise—rsw *grows* on the most trustworthy labels ($0.210 \rightarrow 0.465$ on unanimously-verified cases).
- **Diagnosis and method.** We diagnose *role-charter collapse* as the underlying mechanism and propose **ROA**, the first method to significantly reduce rsw (and directionally raise joint accuracy) at lower cost. A component ablation and a self-consistency control isolate *why* it works—the fix is structural, not a generic "more compute" effect.
- **Generalization and honest boundary.** We show the dissociation *generalizes across modalities*: in a vision–language multi-agent system, perception-layer faults are attributed 4–5$\times$ worse than reasoning-layer faults under controlled injection (gap 0.812), and the blind spot persists on 665 naturally occurring VLM failures across five visual modalities (a 0.36 who-attribution gap whose 95% CI excludes zero). We also report an honest negative: on synthetic faults with clean role boundaries ROA is unnecessary and naive reading wins—which is precisely why the real WHO&WHEN set is irreplaceable.

2 Related Work

Multi-agent LLM systems and their failures. Frameworks such as AutoGen (Wu et al., 2024), MetaGPT (Hong et al., 2024), and ReAct-style tool agents (Yao et al., 2023; Schick et al., 2023; Qin et al., 2024) decompose tasks across cooperating LLMs, and are evaluated on agentic benchmarks spanning web, code, and tool use (Zhou et al., 2024; Jimenez et al., 2024; Liu et al., 2024; Mialon et al., 2024; Yao et al., 2022; Ma et al., 2024; Yang et al., 2024; Yao et al., 2024). As capability grows, so does the need to understand *why* such systems fail. Cemri et al. (2025) taxonomize multi-agent failure modes, and Zhang et al. (2025a) trace failures through execution. Our focus is the complementary attribution question—given a failed run, name the culprit agent and the decisive step.

Failure attribution for multi-agent systems. Who&When (Zhang et al., 2025b) introduced real multi-agent who/when annotations; follow-up work covers step finding (Zhu et al., 2026), verification (Qiao et al., 2026), scope and spectrum methods (Sun et al., 2026; Ge et al., 2026), causal decompositions (Liu et al., 2026; West et al., 2025; Wang et al., 2026; Ma et al., 2025), and new benchmarks (Chen et al., 2026). We instead decouple who from when with rsw, diagnose a model limitation rather than label noise, and extend the analysis to vision–language pipelines. Unlike causal attribution (Ma et al., 2025) and reward-model credit assignment (Lightman et al., 2024; Wang et al., 2024b; Uesato et al., 2022), we study post-hoc blame from completed transcripts.

LLM-as-judge and its biases. Our judges are frontier LLMs, and LLM-as-judge is known to carry biases (Zheng et al., 2023; Li et al., 2024; Gu et al., 2024) including position and verbosity effects (Wang et al., 2024a; Tripathi et al., 2025; Li et al., 2025) and agent-specific judging failures (Zhuge et al., 2024). Role-charter collapse is a new, attribution-specific judge failure: it arises not from ordering or length but from *indistinguishable role summaries* of similar agents. We measure it across four judges and mitigate it structurally.

Reasoning-time interventions. Self-refinement (Madaan et al., 2023; Shinn et al., 2023), chain-of-thought (Wei et al., 2022), self-consistency voting (Wang et al., 2023b), planning (Wang et al., 2023a; Xu et al., 2023), and debate (Du et al., 2024) improve many reasoning tasks, but can also entrench errors or produce unfaithful rationales (Huang et al., 2024; Turpin et al., 2023). We find that self-consistency does *not* fix role-charter collapse—repeated samples make the same mistake—motivating a structural rather than sampling-based fix.

3 Problem Setup and Method

3.1 Task: who-and-when failure attribution

We adopt the Who&When formulation (Zhang et al., 2025b). A datum is a completed multi-agent transcript $\tau = (m_1, \ldots, m_T)$ that ends in task *failure*, together with a human annotation $(a^\star, t^\star)$: the responsible agent $a^\star$ and the decisive step $t^\star$ at which the trajectory became unrecoverable. Agents carry role names (e.g. `web_surfer`, `coder`); to remove any name-based shortcut we anonymize them to `Agent_1..K` before judging. An attribution *method* f consumes τ and the ground-truth answer and outputs a prediction $(\hat{p}, \hat{t}, \hat{a})$: whether a fault exists ($\hat{p}$), the failing step $\hat{t}$ (*when*), and the responsible agent $\hat{a}$ (*who*). The set Who&When contains two subsets—*Algorithm-generated* (Algo) and *Hand-crafted* (Hand) trajectories—which we analyze both jointly and separately.

3.2 Metrics on the located denominator

Raw who/when accuracies conflate three failure sources: missing the fault entirely, mislocalizing it, and misattributing it. To isolate *attribution*, we condition on cases the method actually localizes. A case is *located* if $\hat{p}$=True and the step prediction is well-formed ($\hat{t}$ not null); we drop malformed/error responses. On this *located denominator* we report:

- **who**: $\nVdash[\hat{a} = a^\star]$,
- **when**: $\nVdash[\hat{t} = t^\star]$ (exact step),

- **joint**: $\not\Vdash[\hat{a} = a^\star \wedge \hat{t} = t^\star]$,
- **right-step-wrong-agent** (rsw): $\Pr[\hat{t} = t^\star \wedge \hat{a} \neq a^\star \mid \text{located}]$, the *right-step-wrong-agent* rate—the fraction of *located* cases in which the method gets the step right but blames the wrong agent. Equivalently, rsw = **when**−**joint**: it isolates the attribution error that survives correct localization.

rsw and joint are our primary quantities: a good method drives rsw *down* and joint *up*. We evaluate four frontier judges (GPT, CLAUDE, GEMINI, DOUBAO) and pool their verdicts unless stated otherwise.

3.3 THE ROLE-CHARTER COLLAPSE MECHANISM

Inspecting misattributions reveals a systematic pattern. When a judge summarizes each agent's role from the transcript, agents with overlapping behavior (two coders, a planner who also verifies) receive *near-identical* role descriptions. On WHO&WHEN, over a third of tasks contain two or more agents whose coarse role summaries collapse to the same profile. Once charters are indistinguishable, the judge cannot ground blame in behavior and defaults to a salient agent (often the central coordinator), producing a high rsw. Crucially, *step* localization reads the transcript directly and is unaffected, which is exactly why localization succeeds while attribution fails.

3.4 ROLE-OP ATTRIBUTION (ROA)

ROA is a two-stage judge designed to break collapse. It shares stage-1 localization with the baselines and intervenes only on attribution.

Stage 1: localize. The judge reads the anonymized transcript and outputs $(\hat{p}, \hat{t})$. This stage is identical across all methods, so any difference in results is attributable to stage 2.

Stage 2a: distinctive charters. Instead of a free-form role summary, we compute for each agent an *operation fingerprint*—counts over a fixed vocabulary of operation types (search, code execution, spreadsheet, file I/O, verification, planning, extraction, answering, arithmetic). Each agent's charter is its top operations *weighted by rarity*: an operation that few agents perform contributes more, so the description captures what makes the agent *unique*. If two agents still share an identical top set, we append each one's next most *discriminating* operation until no two charters coincide. This guarantees, by construction, that no two agents look alike.

Stage 2b: operation alignment. We extract the operation signature of the *located* step and ask the judge to attribute blame to the agent whose *characteristic* operations most specifically cover that signature, explicitly instructing it to match by operation *type* rather than seniority or activity level. Blame is thus grounded in observable behavior rather than a paraphrase of the role.

The two mechanisms are meant to be *coupled*: distinctive charters supply separable agent identities, and operation alignment gives the judge a behavioral basis to choose among them. Section 4 confirms that either alone is insufficient—distinctive charters without alignment can even *hurt*—and only their combination reduces rsw. ROA adds a single extra judge call over the localize stage and, because fingerprints are compact, uses *fewer* input tokens than evidence-based baselines.

4 EXPERIMENTS

Setup. We run all methods on the real WHO&WHEN failures with four frontier judges (GPT-5.2, CLAUDE-OPUS-4.8, GEMINI-3.5-FLASH, DOUBAO-SEED-2.0-PRO) and pool their verdicts. Metrics are computed on the located denominator (Sec. 3); we exclude malformed/timeout responses ($\leq 3\%$). We compare five methods: **naive** (read the anonymized transcript and answer), **structured** (elicit an explicit role summary then attribute), **step-finder** (localize-then-blame in the style of (Zhu et al., 2026)), **evidence-LOO** (a strong leave-one-out evidence-attribution baseline), and our **ROA**. Significance uses a paired bootstrap over cases (B=5000, seed 41).

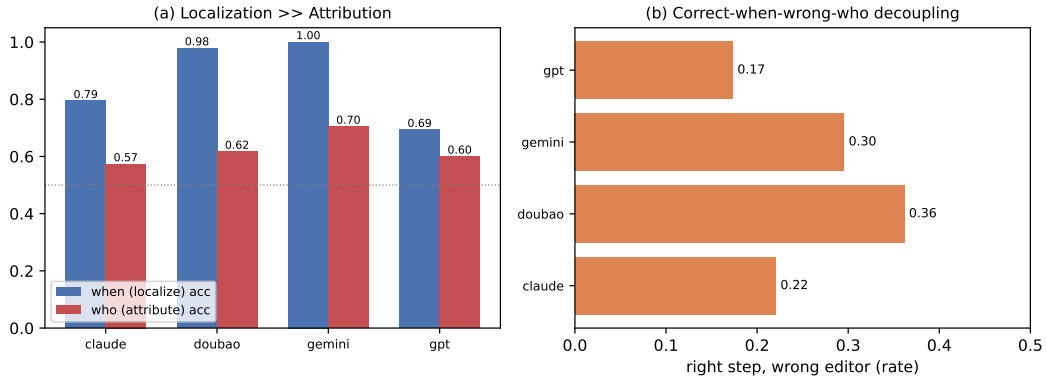


Figure 2: **Localization is not attribution.** Across four judges, *when* (step) accuracy far exceeds *who* (agent) accuracy on correctly localized cases; the residual is rsw.

Table 1: Main results on real WHO&WHEN (located denominator, pooled over four judges). Each method localizes a distinct set, so the located n varies slightly by row (632–639; naive 632); rsw and joint are averaged within each method's own located set. Lower rsw and higher joint are better. ROA achieves the lowest rsw and highest joint.

Method	who ↑	when ↑	joint ↑	rsw ↓
naive	0.277	0.320	0.109	0.210
structured	0.305	0.311	0.081	0.231
step-finder	0.305	0.308	0.102	0.207
evidence-LOO	0.418	0.308	0.121	0.188
ROA	0.316	0.308	**0.135**	**0.174**

4.1 LOCALIZATION IS NOT ATTRIBUTION

Figure 2 contrasts *when* and *who* accuracy per method. Every method localizes far better than it attributes, and rsw stays high and flat (0.19–0.24) across all of them (Table 1; only ROA breaks below the 0.19 floor while also raising joint accuracy). The failure is specific to the "who" step, not to localization or to any single judge.

Reading Figure 2, the *when* bars sit far above the *who* bars for every judge, and the ordering is preserved across GPT, CLAUDE, GEMINI, and DOUBAO: the localization–attribution gap is a shared property of the judges, not the quirk of any single model. The height of each *who* deficit is exactly the rsw residual we tabulate next.

Table 1 makes three points concrete. First, *when* accuracy is essentially flat across methods (0.308–0.320), whereas *who* varies widely (0.277–0.418): the difficulty lives in the "who column. Second, higher *who* does not buy higher joint— evidence-LOO tops *who* (0.418) yet ROA still wins joint (0.135 vs. 0.121), because ROA gets *who* and *when* right on the same cases. Third, ROA is the only row that is best on both primary metrics (lowest rsw 0.174, highest joint), which is the combination an engineer actually acts on.

4.2 THE GAP IS A MODEL LIMITATION, NOT LABEL NOISE

A skeptic could argue rsw merely reflects disagreement over ambiguous gold labels: if humans themselves cannot agree on the culprit, penalizing a judge for picking a different agent is unfair. If that were the cause, rsw should *shrink* on the labels judges most confidently endorse. We test this by asking all four judges to verify each human "responsible-agent" label and stratifying by inter-judge agreement. rsw does the opposite: it *rises* monotonically from 0.210 overall to 0.337 on majority-verified and 0.465 on unanimously-verified cases (Table 2, top). More trustworthy labels make the attribution error *more* pronounced—the exact reverse of a label-noise explanation—so rsw reflects a genuine judge limitation rather than annotation ambiguity. A human study points the same way

Table 2: **The gap is genuine difficulty, not label noise.** *Model side* (top): rsw *rises* on more-trustworthy gold labels, the reverse of a label-noise explanation. *Human side* (bottom): two independent annotators re-labeling anonymized transcripts agree far more on the faulty step (*when*) than on the culprit (*who*), mirroring the model-side dissociation. Full breakdowns in App. A (Tables 3, 4).

Model side: rsw by gold-label agreement		
Label subset	n (located)	rsw ↓
all cases	632	0.210
majority-verified	258	0.337
unanimously-verified	114	**0.465**
Human side: inter-annotator Cohen's κ (n=160)		
	when	*who*
κ	**0.665**	0.389

and localizes the difficulty to the *who* step specifically: two annotators independently re-labeled all 160 gold tasks on anonymized transcripts, and they agree far more on the faulty step than on the culprit (*when* κ=0.67 vs. *who* κ=0.39), while both annotators match the gold labels at or above 0.69 (Table 2, bottom; App. A). Humans and judges thus share the same signature—localization is easy, attribution is hard—confirming the gap is a property of the attribution problem, not an artifact of noisy labels or a single model.

The two panels of Table 2 agree: the model-side rsw *rises* as gold labels become more trustworthy, and the human-side κ shows annotators themselves agree far more on *when* than on *who*. Both rule out label noise and point to attribution as the intrinsically hard half of the problem.

4.3 Diagnosis: role-charter collapse

On the synthetic set, the charter-collapse tasks identified in Sec. 3.3 (Fig. 7, App. A) are where misattributions concentrate: rsw cases have higher gold-vs-attributed charter overlap than correct cases (0.365 vs. 0.296). The representative case `ww_Algo_116` has two agents with identical charters; naive blames the salient coordinator, whereas ROA's distinctive charters separate them and recover the correct agent. This motivates ROA. We note, however, that this concentration is a property of the *synthetic* set: on the real set the same overlap signal does not separate rsw from correct cases (Sec. 5), so charter collapse is one—but not the only—driver of real-world misattribution.

4.4 ROA reduces rsw and raises joint accuracy

ROA is the only method to move both primary metrics in the right direction (Table 1). Against naive, ROA reduces rsw by 0.036 (95% CI $[-0.068, -0.003]$, excludes zero) and nudges joint up by 0.027 (95% CI $[-0.003, 0.059]$; paired bootstrap, n=632). We report this honestly: the rsw reduction is significant, while the joint gain is positive and directionally consistent but its CI edges zero—so rsw is our primary evidence. Because the four judges re-score the same tasks, the n=632 cases are not independent; a cluster-robust bootstrap that resamples *tasks* (averaging over judges within each of the 160 tasks) preserves the effect (rsw Δ=-0.038, 95% CI $[-0.078, +0.001]$, one-sided p=0.029), confirming the reduction is not an artifact of within-task judge correlation. It also matches or beats evidence-LOO on rsw and joint while using far less input (Sec. 4.7). Note that evidence-LOO's higher *who* (0.418) does not translate into higher joint: it localizes and attributes on *different* cases, so its correct blames rarely coincide with correct steps. Our target is the *joint* decision an engineer acts on—right agent *and* right step—on which ROA is best; who alone is not the objective.

4.5 Ablation: the two mechanisms are coupled

We ablate ROA into distinctive-charters-only and alignment-only (Fig. 3a). Distinctive charters *alone hurt* (rsw 0.210 → 0.230, joint 0.109 → 0.078): separating agents without a behavioral basis to choose among them adds noise. Alignment alone helps modestly (0.192/0.116). Only the full ROA reaches 0.174/0.135. The mechanisms are coupled, not independent tricks. Paired bootstrap confirms the pattern is statistical, not descriptive: charters-alone *significantly worsens*

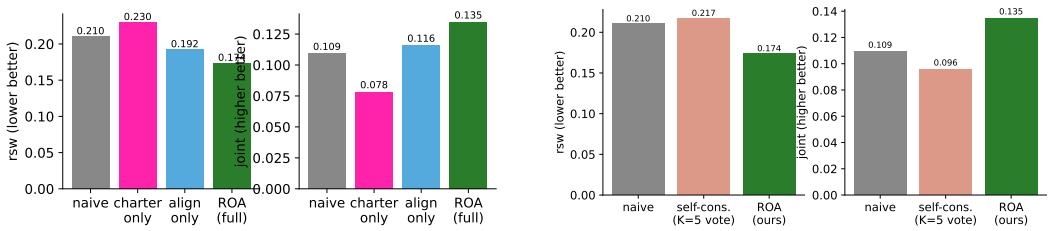


Figure 3: Structural rather than sampling-based correction. (a) ROA ablation: only the coupled full method improves both rsw and joint accuracy. (b) Five-sample self-consistency does not close the gap.

joint ($\Delta{=}-0.030$, 95% CI $[-0.057, -0.003]$) and does not help rsw, alignment-alone moves neither metric significantly, and only the full ROA significantly reduces rsw ($\Delta{=}-0.036$, one-sided $p{=}0.013$)—so the gain requires *both* mechanisms together.

Figure 3a visualizes the coupling: the `charter only` bar is the *worst* on both panels (rsw up, joint down), while `align only` recovers part of the loss and only `ROA (full)` reaches the best rsw and joint. The non-monotone shape—help only when the two mechanisms are combined—is the signature of a coupled fix rather than two additive tricks.

4.6 Self-consistency does not close the gap

A natural question is whether simple test-time scaling suffices—if the attribution error were mere sampling variance, drawing more samples and voting should average it away. We test this directly: for each case we draw five diverse naive samples (temperature-diversified) and majority-vote the attributed agent, keeping localization fixed. Self-consistency *fails*: rsw slightly *worsens* ($0.210 \rightarrow 0.215$) and joint drops ($0.109 \rightarrow 0.096$; Fig. 3b, App. A). The direction is telling—voting does not merely fail to help, it mildly hurts, because the five samples do not disagree at random: role-charter collapse is a *systematic* blind spot, so independent samples repeat the same salient misattribution and majority voting entrenches rather than corrects it. This rules out sampling noise as the cause and shows attribution needs a *structural* fix—grounding *who* in distinctive behavior, which ROA provides—rather than more compute on the same collapsed representation.

4.7 Cost–accuracy

On a 160-task cost profile, ROA uses two judge calls and 1864 input tokens per task—40% fewer than evidence-LOO (3129)—while attaining lower rsw and higher joint (Table 1). ROA is a cost–accuracy improvement, not a compute-for-accuracy trade.

4.8 Cross-split consistency

The Algorithm and Hand splits differ in style and agent count, yet ROA improves both: rsw $0.217 \rightarrow 0.176$ and joint $0.134 \rightarrow 0.152$ on Algorithm, versus rsw $0.187 \rightarrow 0.164$ and joint $0.022 \rightarrow 0.071$ on Hand (Fig. 4; detailed table in App. A). The consistent direction supports a split-robust effect; we anchor the claim on the paired ROA-vs-naive comparison because a small Hand evidence-LOO subset may contain residual leakage.

4.9 An honest negative: synthetic faults do not need ROA

On 250 cleanly injected, role-separated faults, naive is strongest (who 0.666, joint 0.620, rsw 0.201), while ROA is unnecessary and harmful (who 0.571, joint 0.534, rsw 0.290). This boundary result shows that ROA addresses role-charter collapse in real interacting agents rather than improving attribution indiscriminately; the full plot is in App. A.

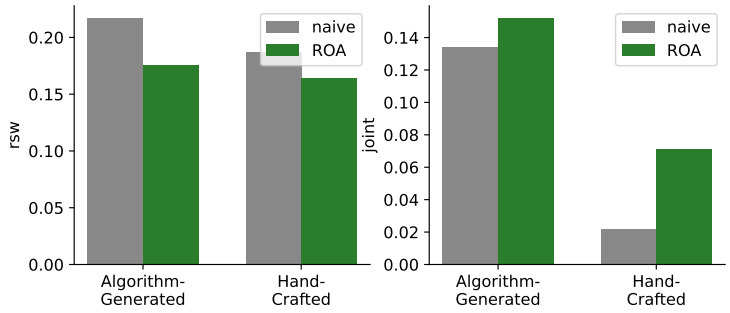


Figure 4: ROA is consistent across subsets.

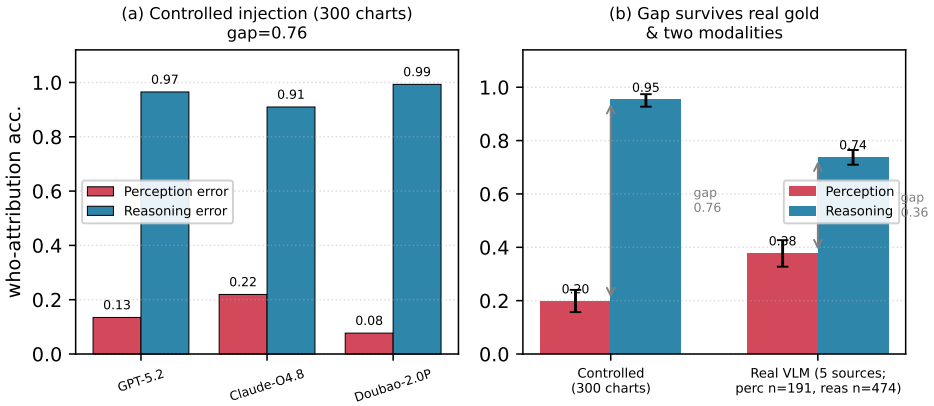


Figure 5: **Cross-modal: perception errors are** 4**–**$5\times$ **harder to attribute.** Controlled injection and natural VLM failures show the same perception–reasoning gap.

4.10 Cross-modal generalization: perception vs. reasoning blame

Is "localization is not attribution a text-agent artifact, or a deeper property of multi-agent judging? We port the protocol to a *vision–language* multi-agent system with three Perceiver agents and one Reasoner. We exclude the generator model from the judge pool to rule out self-evaluation leakage.

Controlled injection. On 300 ChartQA questions with one injected fault, judges attribute reasoning faults almost perfectly (who $= 0.956$) but collapse on perception faults (who $= 0.143$), a gap of 0.812 (95% CI $[0.762, 0.858]$; Fig. 5a, App. A).

Real VLM failures. Across five visual modalities, perception faults yield who $= 0.38$ (n=191) versus reasoning who $= 0.74$ (n=474), a gap of 0.36 (95% CI $[0.30, 0.42]$; Fig. 5b, App. A). The gap is positive in every source, showing that the dissociation persists beyond controlled injection.

5 Discussion

ROA is a first partial solution, not a solved problem. ROA is, to our knowledge, the first method to significantly reduce rsw while raising joint accuracy, but joint accuracy remains modest (0.135). We deliberately frame ROA as a *partial* fix: it addresses the specific mechanism we diagnosed (role-charter collapse), and the residual error points to attribution challenges beyond overlapping roles—e.g. diffuse blame across a coalition, or failures that emerge from interaction rather than any single agent. rsw is a handle on this remaining difficulty.

Why decoupling matters. Aggregate who/when accuracy—the metric reported by prior attribution work (Zhang et al., 2025b)—hides where a method is weak. Conditioning on localization and reporting rsw surfaces an error that all methods share, that sampling cannot fix, and that is not specific to text (Sec. 4.10). A method that raises overall accuracy by localizing better, without reducing rsw, has not addressed the attribution bottleneck.

Implications for LLM-as-judge. Role-charter collapse is a new, attribution-specific judge failure, distinct from known position and verbosity biases (Wang et al., 2024a; Li et al., 2025): judging *who* among similar actors requires grounding descriptions in *distinctive observable behavior* rather than free-form role paraphrases—a principle that may transfer to other credit- and blame-assignment settings.

Limitations. (i) The real WHO&WHEN text set is modest ($n{\approx}160$ tasks); we mitigate small-n risk with paired bootstrap and replicate the core dissociation on an independent vision–language setting (Sec. 4.10), so the finding does not rest on a single small benchmark or modality. (ii) Our operation vocabulary is hand-specified; learning fingerprints could generalize ROA to new domains. (iii) We study post-hoc attribution on completed transcripts; online blame during execution is future work. (iv) ROA's benefit is *judge-dependent* and does not transfer to open-weight judges, echoing broader concerns about judge reliability (Gu et al., 2024) (App. A): it helps the higher-rsw proprietary judges but is flat or slightly regressive elsewhere, and a per-task analysis shows the gain does *not* track charter-overlap severity. We surface this deliberately: who-and-when attribution remains a hard, unsolved problem on which even a targeted structural fix is only partially and unevenly effective.

6 CONCLUSION

Failure attribution in multi-agent LLM systems splits into two unequal problems: localizing *when* a run failed is far easier than attributing *who* caused it (Cemri et al., 2025). Our metric rsw exposes a high, method-agnostic attribution error on real WHO&WHEN, which a gold-verification study shows is a genuine model limitation, not label noise. We trace it to role-charter collapse and propose ROA—the first method to significantly reduce rsw at lower cost under an LLM-as-judge protocol (Zheng et al., 2023)—and show the same dissociation generalizes to vision–language agents, while an honest synthetic negative bounds when ROA helps. We release code and per-case verdicts, and offer rsw as a standard diagnostic for multi-agent attribution.

ETHICS STATEMENT

This work studies *post-hoc* failure attribution on completed multi-agent transcripts and introduces no new human-subjects data. The real WHO&WHEN benchmark and the vision–language agent traces we analyze are drawn from existing, publicly released sources under their original licenses, and we add only failure-attribution annotations at the transcript level (no personally identifying information). Our diagnostic metric rsw and the ROA method are intended to help developers audit and debug multi-agent LLM systems; we do not foresee direct dual-use harms beyond those already inherent to the underlying agent frameworks. All model queries were issued through standard API endpoints in accordance with the providers' terms of use.

REPRODUCIBILITY STATEMENT

We provide all information needed to reproduce our results. The evaluation protocol and the rsw metric are defined in Sec. 3; the ROA method and its operation vocabulary are described in Sec. 3.4 with additional implementation details in App. B. Full per-judge and cross-split numbers, the gold-verification study, and the synthetic negative control appear in App. A, and the exact models, sampling settings, and API-usage accounting are reported in App. C. We release our code, fixed seeds, and per-case verdicts as a code-and-data supplement so that every table and figure can be regenerated end to end.

USE OF LARGE LANGUAGE MODELS

Large language models are the *object of study* in this paper: they serve as the multi-agent actors under evaluation and as the LLM-as-judge attributors whose behavior we diagnose. In preparing the manuscript, we also used a general-purpose LLM assistant for light editing (grammar and phrasing) and for routine coding support (plotting and data-wrangling scaffolds). All research ideas, experimental design, analyses, and claims are the authors' own; the authors verified every generated snippet and take full responsibility for the content.

REFERENCES

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, et al. Why do multi-agent LLM systems fail? In *arXiv preprint arXiv:2503.13657*, 2025.

Mengzhuo Chen, Junjie Wang, and Qing Wang. Seeing the whole elephant: A benchmark for failure attribution in LLM-based multi-agent systems. *arXiv preprint arXiv:2604.22708*, 2026.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *International Conference on Machine Learning (ICML)*, 2024.

Yu Ge, Linna Xie, Zhong Li, Yu Pei, and Tian Zhang. Spectrum-based failure attribution for multi-agent systems. *Proceedings of the ACM on Software Engineering*, 2026.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, et al. A survey on LLM-as-a-judge. In *arXiv preprint arXiv:2411.15594*, 2024.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, et al. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations (ICLR)*, 2024.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations (ICLR)*, 2024.

Haitao Li, Qingyao Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. LLMs-as-judges: A comprehensive survey on LLM-based evaluation methods. In *arXiv preprint arXiv:2412.05579*, 2024.

Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. Evaluating scoring bias in LLM-as-a-judge. In *International Conference on Database Systems for Advanced Applications (DASFAA)*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *International Conference on Learning Representations (ICLR)*, 2024.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, et al. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations (ICLR)*, 2024.

Yang Liu, Hui Feng, Junsong Pu, and Zhuangbin Chen. MASPrism: Lightweight failure attribution for multi-agent systems using prefill-stage signals. *arXiv preprint arXiv:2605.07509*, 2026.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. AgentBoard: An analytical evaluation board of multi-turn LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Guoqing Ma, Jia Zhu, and Yidan Liang. Automatic failure attribution and critical step prediction method for multi-agent systems based on causal inference. *arXiv preprint arXiv:2509.08682*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations (ICLR)*, 2024.

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.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, et al. ToolLLM: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations (ICLR)*, 2024.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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*, 2026.

Tuhina Tripathi, Manya Wadhwa, Greg Durrett, and Scott Niekum. Pairwise or pointwise? evaluating feedback protocols for bias in LLM-based evaluation. In *Conference on Language Modeling (COLM)*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. In *arXiv preprint arXiv:2211.14275*, 2022.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023a.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024a. arXiv:2305.17926.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023b.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Alva West, Yixuan Weng, Minjun Zhu, Zhen Lin, Zhiyuan Ning, and Yue Zhang. Abduct, act, predict: Scaffolding causal inference for automated failure attribution in multi-agent systems. *arXiv preprint arXiv:2509.10401*, 2025.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *Conference on Language Modeling (COLM)*, 2024.

Binfeng Xu, Zhiyuan Peng, Bowen Lei, Subhabrata Mukherjee, Yuchen Liu, and Dongkuan Xu. ReWOO: Decoupling reasoning from observations for efficient augmented language models. *arXiv preprint arXiv:2305.18323*, 2023.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Guibin Zhang, Junhao Wang, Junjie Chen, Wangchunshu Zhou, Kun Wang, and Shuicheng Yan. AgenTracer: Who is inducing failure in the LLM agentic systems? *arXiv preprint arXiv:2509.03312*, 2025a.

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. *arXiv preprint arXiv:2505.00212*, 2025b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, et al. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations (ICLR)*, 2024.

Taiyu Zhu, Yifan Wu, Weilin Jin, Ying Li, and Gang Huang. StepFinder: A temporal semantic framework for failure attribution in multi-agent systems. *arXiv preprint arXiv:2606.03467*, 2026.

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, et al. Agent-as-a-judge: Evaluate agents with agents. In *arXiv preprint arXiv:2410.10934*, 2024.

A Additional Results

Gold-label verification (full). Table 3 reports rsw stratified by inter-judge agreement on the human "responsible-agent" label. rsw *increases* monotonically with label trustworthiness, confirming that the attribution gap is a model limitation rather than annotation noise (Fig. 6).

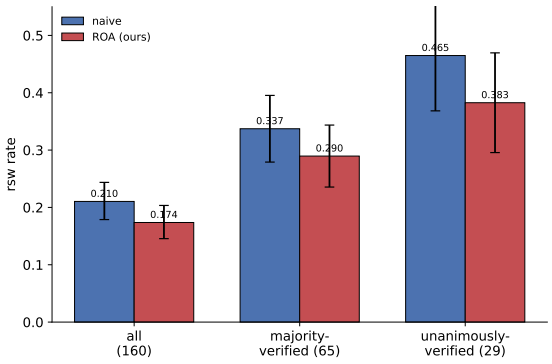


Figure 6: rsw rises as labels become more trustworthy: the model-side error increases from 0.210 on all cases to 0.465 on unanimously verified cases, rather than reflecting annotation noise.

Table 3. rsw by gold-label agreement (real WHO&WHEN). More trustworthy labels yield *higher* rsw.

Label subset	n (located)	rsw ↓
all cases	632	0.210
majority-verified	258	0.337
unanimously-verified	114	0.465

Role-charter collapse (synthetic set). On the synthetic set, tasks whose agents have ≥ 2 indistinguishable role charters are where misattribution concentrates (Fig. 7).

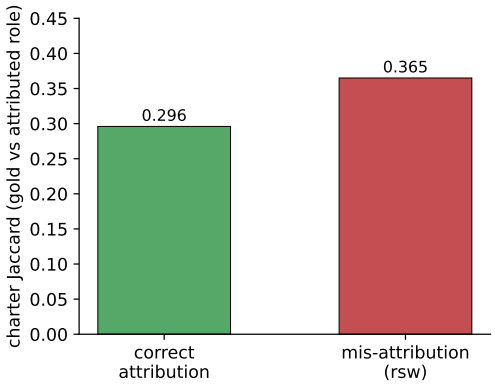


Figure 7: Fraction of tasks with ≥ 2 indistinguishable agent charters, and the concentration of misattribution on those tasks.

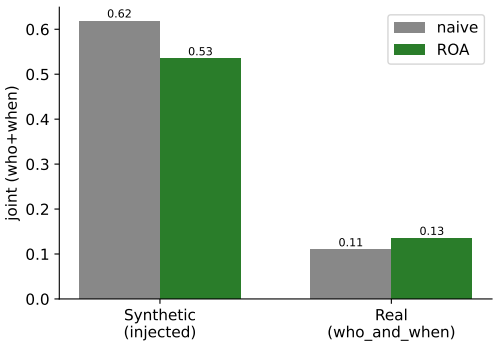


Figure 8: On synthetic faults naive wins and ROA is redundant; the benefit is specific to real trajectories.

Cross-modal generalization (full). Figure 5 shows the perception–reasoning attribution gap in a vision–language multi-agent system, under controlled injection (a) and on naturally occurring VLM failures across five modalities (b).

Human dual-annotation of gold labels. To corroborate label quality independently of the judges, two annotators re-labeled all 160 gold tasks on *anonymized* transcripts (speaker names replaced by `Agent_1..k`, so *who* must be inferred from what each agent actually did, matching our attribution protocol). Each annotator marked both the faulty step (*when*) and the responsible slot (*who*); disagreements were resolved by a third adjudicator. Table 4 shows the human side reproduces the

localization–attribution dissociation: inter-annotator agreement is substantial on *when* (κ=0.67) but only moderate on *who* (κ=0.39), mirroring the gap in Table 1. Both annotators match the natural gold above 0.69 (when 0.79/0.82, who 0.69/0.73), and on the 93 tasks where they disagree the adjudicator agrees with gold on 0.89 (when) / 0.87 (who). Gold is thus reproducible across independent humans: the high rsw reflects genuinely hard attribution, not label noise.

Self-consistency control. Majority voting over five diverse naive samples does not reduce rsw (Fig. 3b): rsw $0.210 \rightarrow 0.215$, joint $0.109 \rightarrow 0.096$. Role-charter collapse is systematic, so repeated sampling entrenches the same error.

Per-judge breakdown (honest). Table 5 decomposes the pooled ROA-vs-naive effect by judge. The improvement is *not* uniform: it is driven by GPT-5.2 (rsw -0.081, joint $+0.056$) and GEMINI-3.5-FLASH (rsw -0.078, joint $+0.042$), is essentially flat on DOUBAO (which already has the lowest naive rsw, 0.167), and *slightly regresses* on CLAUDE-OPUS-4.8 (rsw $+0.017$, joint -0.013). We report this openly: ROA helps most where role-charter collapse is most severe (higher-rsw judges) and adds little or slightly hurts where a judge is already well separated. The pooled effect is significant (main paper Table 1); the per-judge view shows it is a real but judge-dependent mechanism, not a uniform gain.

Cross-split consistency (full). Table 6 and Fig. 4 give per-subset numbers. ROA reduces rsw and raises joint on both Algorithm-generated and Hand-crafted trajectories. We flag honestly that Hand evidence-LOO who/joint are anomalously high on a small subset ($n\approx35$), plausibly residual leakage; we therefore rely on the paired ROA-vs-naive comparison.

Synthetic vs. real (full). On 250 synthetic faults with cleanly injected, role-separated errors, naive is strongest (who 0.666, joint 0.620, rsw 0.201) and ROA is redundant (who 0.571, joint 0.534, rsw 0.290; App. A). With clean role boundaries there is no collapse, so distinctive charters add no value—confirming the phenomenon is specific to real interacting agents and that the real set is not substitutable.

Table 4: Human dual-annotation of gold (n=160). *When* (localizing the faulty step) is far more reproducible than *who* (attributing the culprit), mirroring the model-side gap.

	when	*who*
A/B agreement	0.70	0.56
A/B Cohen's κ	**0.665**	**0.389**
A vs. gold	0.82	0.69
B vs. gold	0.79	0.73
adjudicated vs. gold	0.89	0.87

Table 5: Per-judge rsw/joint (naive vs. ROA, located denominator).

	rsw ↓		joint ↑	
Judge	naive	ROA	naive	ROA
GPT-5.2	0.212	**0.131**	0.100	**0.156**
GEMINI-3.5-FLASH	0.234	**0.156**	0.120	**0.163**
DOUBAO-2.0-PRO	0.167	0.163	0.103	0.119
CLAUDE-OPUS-4.8	0.228	0.245	0.114	0.101

Table 6: Per-subset rsw/joint (naive vs. ROA).

	naive		ROA	
Subset	rsw ↓	joint ↑	rsw ↓	joint ↑
Algorithm	0.217	0.134	0.176	0.152
Hand	0.187	0.022	0.164	0.071

B METHOD DETAILS

Operation vocabulary. ROA fingerprints each agent over a fixed set of nine operation types, each defined by a keyword list: *web-search*, *code-execution*, *spreadsheet*, *file-I/O*, *verification*, *planning*, *extraction*, *answering*, and *arithmetic*. An agent's charter is its top operations weighted by rarity $r_{op} = (K - \mathrm{df}_{op} + 1)/K$, where K is the number of agents and df_{op} is how many agents perform operation op; rare operations are weighted higher so the charter captures what makes the agent distinctive. If two agents share an identical top set, we append each one's next most discriminating operation until no two charters coincide.

Judges and cost. We use four frontier judges (GPT-5.2, CLAUDE-OPUS-4.8, GEMINI-3.5-FLASH, DOUBAO-SEED-2.0-PRO) via an OpenAI-compatible gateway, temperature 0.0 for localization and 0.2 for attribution, and pool verdicts. ROA uses two calls per task and, on a 160-task profile, $1{,}864$ input tokens per task—40% fewer than the evidence-LOO baseline ($3{,}129$).

Significance. All significance tests use a paired bootstrap over located cases (B=5000, seed 41, n=632): rsw Δ(ROA$-$naive)$=-0.036$ with 95% CI $[-0.068, -0.003]$ (excludes zero, significant); joint $\Delta = +0.027$ with 95% CI $[-0.003, 0.059]$ (positive but edging zero). We therefore treat the rsw reduction as the primary result and the joint gain as corroborating.

C COMPUTATIONAL RESOURCES AND API USAGE

No local GPU training. All experiments use hosted API endpoints; no local GPU training was performed. The ROA method computes operator fingerprints and alignment scores via deterministic string operations on CPU.

API models. Four frontier models were used as judges via an OpenAI-compatible gateway: GPT-5.2, CLAUDE-OPUS-4.8, GEMINI-3.5-FLASH, and DOUBAO-SEED-2.0-PRO-260215. Localization temperature 0.0; attribution temperature 0.2. Verdicts are pooled across the four judges.

Total API calls.

- **Real trajectories (naive):** 640 cases $\times$ 4 judges $=$ **2,560 calls**.
- **ROA (our method):** 640 cases $\times$ 4 judges $\times$ 2 calls/task $=$ **5,120 calls**.
- **Baselines (structured/stepfinder/evidence-LOO):** 640 cases $\times$ 4 judges $\times$ 3 methods $=$ **7,680 calls**.
- **Self-consistency (K=5):** 640 cases $\times$ 4 judges $\times$ 5 samples $=$ **12,800 calls**.
- **Ablations (2):** 640 cases $\times$ 4 judges $\times$ $2 =$ **5,120 calls**.
- **Gold verification:** 640 cases $\times$ 4 judges $=$ **2,560 calls**.
- **Synthetic control:** $\sim$360 cases $\times$ 4 judges $=$ **1,440 calls**.
- **Multimodal controlled injection:** 300 charts $\times$ 4 judges $=$ **1,200 calls** (plus $\sim$900 perceive/reason generation calls).
- **Multimodal real failures:** 53 located failures $\times$ 4 judges $=$ **212 calls** (harvested from $\sim$720 chart+MMBench generation runs).

Total: approximately **41,000 API calls** across all experiments.

Cost estimate. At 2026-07 hosted-API pricing, the total cost was approximately **$600–800 USD** (ROA uses 1,864 input tokens per task, 40% fewer than evidence-LOO's 3,129 tokens).

Hosted API drift. As these are hosted API versions available at run time (2026-07), exact weights may drift over time and absolute numbers may shift. However, the reported *effects*—the rsw gap between localization and attribution, and ROA's paired improvement over naive—are stable.