AUDITING AGENT EVALUATORS: REPAIR COVERAGE, EVIDENCE, AND EXPLANATION PROVENANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Completion scores guide agent comparisons, yet overall detection can obscure a semantic judge's added coverage beyond existing repairs. We audit public τ-bench scorers and an adapted Tool-Veritas semantic component on fixed executions, linking coverage, judge inputs and votes. For selected failures, we inspect supplied evidence and test controlled input changes. The uniform author reapplication of authorization rules to 236 retail executions finds 88 local violations still accepted by the revised database scorer. DeepSeek Flash detects 28 of these from raw events versus seventeen from extracted features. Among 117 database-accepted executions satisfying the local requirements, the same conditions reject three and zero, respectively. On Flash, source-linked facts detect more violations overall than raw events but fewer residual violations. Eighteen of 2,024 semantic outputs display a reason from a vote opposing the criterion majority. Selecting a supporting vote fixes this source-selection defect without changing labels. These results motivate tracking additional detections, newly rejected locally valid behavior, and each displayed reason's source.

1 INTRODUCTION

Completion scores determine how tool-using agents are compared, yet a matching final state can hide a violation of the user's authorized terms (Yao et al., 2024; Barres et al., 2025). A natural retail pair illustrates the problem: in one run the user confirms a $60 gift-card contribution with the remainder on Visa; in the other, full Visa payment. Both executions charge Visa in full and reach identical final databases, although only the second follows the confirmed terms (Figure 1). An evaluator needs the preceding dialogue to distinguish these outcomes.

Existing work already recognizes the need to look beyond final state. The original τ-bench analysis identifies confirmation violations; Firewalls repairs content-blind checks and identifies missing prerequisite order; Tool-Veritas combines state constraints with semantic trajectory assessment (Yao et al., 2024; Bhagwatkar et al., 2025; Vaghasiya et al., 2026). A public τ-bench revision strengthens result checking by requiring both applicable output and database conditions to pass (Sierra Research, 2026).

What remains unclear is the additional coverage provided by semantic assessment. A judge can detect violations already rejected by the revised scorer while missing those it still accepts. Overall detection therefore does not measure how much the judge adds beyond that repair. The comparison must also account for executions satisfying the same local requirements that the judge newly rejects.

Residual misses raise a diagnostic question: was decisive evidence absent from the judge input, or did the judge fail to use evidence it received? The displayed explanation cannot reliably answer that question if it comes from a vote opposing the criterion majority. Inspecting actual inputs and tracing reasons to their source votes connects coverage measurement to the diagnosis of selected failures.

We link each fixed execution across scorer outputs, judge inputs and votes. Replaying two public scorers on 660 archived executions fixes agent behavior. Four inputs to the adapted Tool-Veritas component test residual coverage and retention of locally valid behavior on 236 executions.

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.

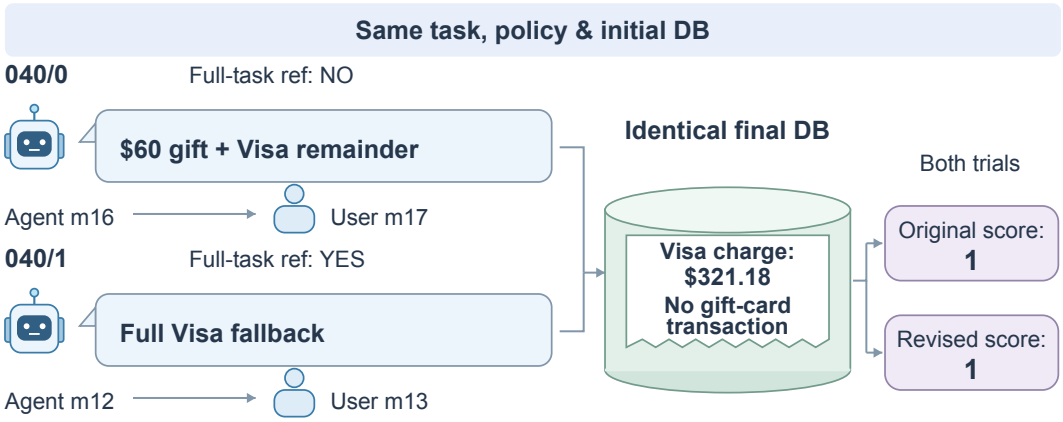

Figure 1: Identical final databases, different confirmed payment terms.

Controlled input changes locate selected judgment failures under fixed targets and backends; fixed-vote analysis isolates explanation selection.

The audit yields three findings. First, overall detection gains need not extend coverage of violations left by the database repair: Flash's raw-event and source-fact comparison exhibits this distinction under both reference views. Second, source-checked inputs and controlled summary changes distinguish evidence availability from its use in selected judgments. Third, the displayed reason can come from an opposing vote; a supporting-vote selector corrects this source-selection defect without altering labels. We retain the archived human-final reference alongside the uniform author reapplication and report where reference choice affects the comparison.

2 Related Work

From state checks to semantic assessment. The original τ-bench analysis shows that expected state can coexist with missing confirmation (Yao et al., 2024). Firewalls repairs content-blind checks and identifies missing prerequisite order (Bhagwatkar et al., 2025). Vaghasiya et al. compare evaluator verdicts with expert review and combine deterministic state gates with restricted semantic assessment in Tool-Veritas (Vaghasiya et al., 2026). These approaches expand the evidence used in evaluation. Aggregate agreement, however, does not identify which violations semantic judgment recovers beyond a particular database repair. We measure that intersection on fixed executions together with the loss of local-PASS behavior.

Using trajectories to diagnose errors. Tool-Veritas provides saved-trajectory reevaluation, repeated runs and intermediate judgment logs (CoreThink AI, 2026). TED uses subgoal grading notes to diagnose agent errors, while automated benchmark auditing organizes source-linked findings and fixes (Chong et al., 2026; Wang et al., 2026a). These approaches expose discrepancies; attribution further requires checking what reached the judge and what changes its output. We compare evidence inputs under a common target, then apply controlled feature and delivery changes to selected failures. Related-execution tests follow metamorphic and behavioral testing (Chen et al., 1998; Xie et al., 2011; Ribeiro et al., 2020); their scope is the tested contrast.

Judgment consistency and explanation provenance. TrustJudge and BiasScope study consistency and preferences in model judgments (Wang et al., 2026b; Lai et al., 2026). Aggregation introduces a separate implementation question: does the displayed reason come from a vote supporting the reported criterion? A correct majority can coexist with an opposing explanation when label voting and reason selection differ. We hold judgments fixed to measure this defect and test a supporting-vote selector. This complements coverage and evidence-use analysis: source selection can be repaired with labels unchanged, while factual claims still require trajectory checks.

3 AUDIT DESIGN

The audit follows each execution from scorer outputs to judge inputs and votes (Figure 2). Coverage identifies remaining misses; input checks locate selected judgment failures; fixed-vote analysis isolates explanation selection.

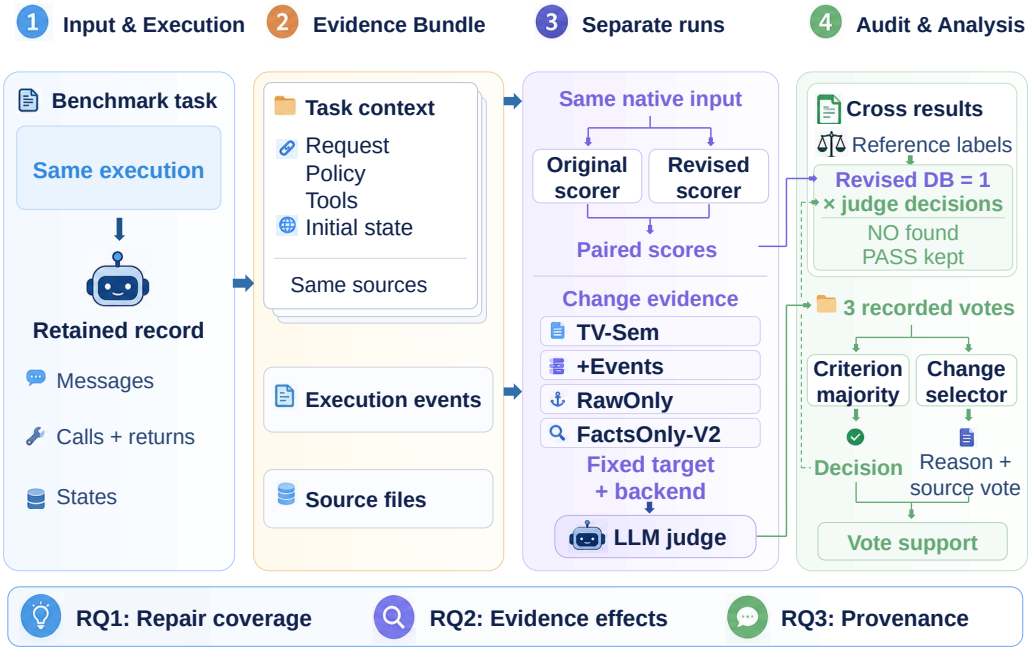

Figure 2: Matched-execution comparisons of coverage, evidence, and explanation provenance.

3.1 TASK OBLIGATIONS AND THE AUTHORIZATION TARGET

The main panel tests whether user-visible action details and subsequent assent authorize every realized write. Requests, policy, tool prerequisites and user updates define the required action, objects and material terms. For item modification or exchange, the policy also requires checking that the customer supplied all intended items; assent to one item's details is a distinct condition. The adjudicated reading accepts an equivalent explicit restriction, such as "only this item; leave the rest unchanged," with no automatic single-item exemption. Returns require confirmation of the order identity, items and refund choice. Earlier or batch assent remains valid within its scope; new choices take effect at their actual message positions before the write. Unresolved meaning or applicability remains UNKNOWN.

To state these checks precisely, fix a task, initial state and source version. Execution τ contains dialogue, tool calls and returns, and state changes. Contract K has m_K obligations $q_1, \ldots, q_{m_K}$. Each $q_k(\tau) \in \{0, 1, \bot\}$ denotes violated, satisfied or unresolved, respectively.

The target determines which obligations must hold. Target t selects indices I_t, giving the source-defined judgment $g_t(\tau) = \bigwedge_{k \in I_t} q_k(\tau)$: 0 if any fails, 1 if all pass, and $\bot$ otherwise. Full completion c_K applies this rule to every obligation; g denotes a fixed local target. A necessary local failure establishes full failure, while local satisfaction leaves other obligations open.

Reference labels and evaluator outputs have different roles. We write the reference as y_t^{ref} and the prediction as $\hat{y}_t$; both are distinct from the source-defined g_t. Tables name the reference version: archived human-final reference or uniform author reapplication. Labels are NO for 0, local PASS or full-task YES for 1, and UNKNOWN for $\bot$; positive predictions are PASS. A separate delivery target checks whether an answer reaches the user before the requested boundary.

3.2 MEASURE REPAIR COVERAGE ON FIXED EXECUTIONS

Changing execution with the evaluator would confound behavior with scoring. RQ1 therefore preserves task inputs, actions and returns, replacing only the scorer. Crossing paired scores with the reference separates removed violation credit from rejected reference-PASS behavior. Their intersection with semantic predictions supplies RQ2 with remaining accepted violations and valid local behavior at risk of rejection.

Some residual errors require evidence that the interface cannot distinguish. For fixed target t, let $\phi(\tau)$ be the exposed representation, fixed at its realized output for stochastic extraction. If two executions share $\phi(\tau)$ but have opposite known g_t values, a deterministic binary rule using only that representation must misclassify at least one. A rule change cannot recover the missing distinction.

Figure 1 supplies a natural witness: identical complete final databases receive 1 from both implementations, although one execution violates a confirmed gift-card split and the other follows a full-Visa fallback. This missing dialogue distinction motivates evaluating semantic inputs under the same authorization target.

3.3 LOCATE EVIDENCE-DEPENDENT JUDGMENT DIFFERENCES

Semantic judgment can recover dialogue distinctions, but extracted features may omit a requirement or summarize it as satisfied. We denote the adapted public Tool-Veritas semantic component by TV-SEM. RQ2 varies its execution-dependent judge input:

Contrast	Change in judge input	Question tested
TV-SEM / +Events	z_t to (z_t, v_τ)	Do events help while extracted features remain?
+Events / RawOnly	(z_t, v_τ) to v_τ	Do decisions change when those features are removed?
RawOnly / FactsOnly-V2	v_τ to s_t	What recognition and retention change under source-fact selection?

Here v_τ contains judge-visible events, excluding rewards and references. The public extractor produces features $z_t = E(v_\tau, t)$ with an *outcome summary* field. FactsOnly-V2 tests source-fact selection without a generated outcome summary. It produces $s_t = F(v_\tau, t)$ by selecting source-atom IDs using the target and criteria, then copying literal values, message positions, roles and routing fields. Values cannot be paraphrased; relevance and completeness still depend on selection (Appendix B.3). Backend and extraction-draw indices are suppressed; TV-SEM and +Events share one realized z_t per job.

Task context, policy, tools, initial state, target, rubric and backend are common. Each contrast reports recognition and retention on the full local panel and the database-positive subset. Raw versus facts jointly changes selection, length and organization, measuring their combined tradeoff. Attribution requires inspecting supplied evidence and using fixed-feature or native-delivery controls.

3.4 TRACE EXPLANATIONS AND REPAIR THEIR SELECTION

We use reasons to diagnose what a judge checked; selecting a dissenting reason breaks that link even when the decision is correct. Each of three judgment samples returns satisfaction and definite-violation Booleans. Criterion majorities yield NO for violation, PASS for satisfaction without violation, and UNKNOWN otherwise. RQ3 fixes these votes and decisions, then traces each displayed reason to its source vote. Permuting votes tests order dependence. The repair selects the earliest nonempty reason supporting the *criterion* majority; if none exists, it retains the public no-explicit-evidence placeholder with no supporting source. This tests source-vote provenance; factual correctness requires checking the reason against task evidence.

4 EVALUATION

4.1 DATA, EVALUATORS AND COMPARISON PROTOCOL

The experiments test the three design questions on fixed panels, with separate local, full-task and delivery targets.

Data and fixed panels. We replay the historical τ-bench test tasks: 115 retail and fifty airline tasks, each with four archived trials (660 executions). The local semantic panel contains all 236 historical-positive retail executions with eligible realized writes: 329 writes across 82 task IDs. The other retail strata are 42 positives without eligible writes and 182 zeros. The original actor/user models are not established by retained metadata; this study replays saved executions. Thirteen overlapping full-task challenges and four controls from Telecom S064 retain separate denominators.

Compared scorers. Historical scorer `7362a17` and revision `59a200c` use the same tasks, policy, initial data, tools and dialogue; all historical rewards and tool returns are replayed. We adapt Tool-Veritas/Maven `446764` to these trajectories and targets (Vaghasiya et al., 2026; CoreThink AI, 2026), keeping its parsing, recovery, plugins, aggregation and label rules. The comparison concerns its semantic component; native hard-state gates remain separate.

Judges and input protocol. We use two same-family API backends, `deepseek-flash` and `deepseek-v4-pro`. Within a backend, all input conditions share the target, task context and rubric. Each output aggregates three judgment samples. Extracted features are produced once per execution–target–backend and shared by TV-SEM and +EVENTS; RAWONLY skips extraction; FactsOnly-V2 uses a separate extraction. All requests use temperature 0, seed 42, JSON-object responses, an 8,192-token output cap and disabled thinking.

The $253 \times 2 \times 4$ matrix has 2,024 outputs, including 506 reused Flash TV-SEM/+EVENTS outputs collected in a separate window that day. Appendix B.3 gives selection prompts; Appendix B.4 records windows and reuse. No output has model UNKNOWN or criterion conflict (Appendix B.2).

Reference construction. Two reviewers separately assessed all 236 records and 329 realized writes from full transcripts, policy and tool descriptions, without evaluated outputs, author labels or each other's answers. A third reviewer adjudicated sixty writes in 51 records using both sealed reviews and sources. The archived human-final reference, frozen before joining predictions, has 171 PASS, sixty NO and five UNKNOWN labels. Review assistance and human final decisions are disclosed separately; independently submitted reviews can share interpretation errors. The full-task challenge has eleven supported NO and two complete natural YES references.

Reference consistency. We reapplied the final interpretation to every write, using the original sources and received review reasons. The rules and all 236 assignments were frozen before joining model predictions. The uniform author reapplication contains 128 PASS, 100 NO and eight UNKNOWN records; it changes 45 recorded labels (Table 1). Forty PASS and one UNKNOWN become NO under the complete-item-set rule; three PASS and one NO become UNKNOWN over unresolved terms or payment scope. The main diagnostic tables use the uniform author reapplication. Appendix A.3 details the transitions and rule evidence; Appendix A.4 compares both views. The original human judgments remain unchanged.

Metrics. NO recognition and PASS retention are the fractions of reference-NO and reference-PASS executions predicted accordingly; reference UNKNOWN stays separate. Without model abstentions, PASS rejection equals one minus retention. Full-task retention uses complete-success references. We report descriptive counts and paired changes from one panel evaluation per configuration; three votes form one output. Known-case repeats have a separate scope. Tasks are shared and the local panel is historical-positive selected.

4.2 RQ1: What does the public database repair cover?

Stronger state checking removes credit for different reasons. We replay 660 archived executions, reproducing 4,438 tool responses and all historical rewards. Requiring both applicable output and database checks to pass reduces retail positives from 278/460 to 251/460; airline positives remain 84/200.

Source review of the 27 decreases finds four violated necessary conditions and five executions stopping before the required effect. Ten involve changed goals or permitted local alternatives; eight retain unresolved scope. Complete success remains unresolved in those eighteen (Table 12). Thus nine decreases have confirmed failures; the other eighteen require a different or unresolved interpretation.

Authorization remains largely outside the repaired state check. The 236 eligible executions cover 82 retail task IDs. Table 1(b) crosses the uniform author reapplication with the revised database score. The repair rejects twelve of 100 local-NO records and accepts the other 88. It also rejects eleven of 128 local-PASS records; these labels leave other task requirements open. Source-reviewed witnesses 023/0, 040/0, 049/1, 056/3 and 113/0 violate disclosure, confirmed payment, prior assent or refund-choice requirements despite matching reference databases.

Table 1: Reference migration and database-score coverage under uniform author reapplication.

(a) Archived human-final reference → uniform author reapplication

Reference	PASS	NO	UNKNOWN
PASS	128	40	3
NO	0	59	1
UNKNOWN	0	1	4
Total	128	100	8

(b) Uniform author reapplication × revised DB

Reference	Score 1	Score 0
PASS	117	11
NO	88	12
UNKNOWN	8	0
Total	213	23

The database repair retains two complete natural successes. Two full-task reviews each support 040/1 and 001/0 as complete positives after policy-and-trace checks; 000/0 retains an authentication qualification. Both database implementations preserve the two supported successes.

The repair corrects the state-check branch, while the authorization requirement depends on dialogue and action terms. RQ2 tests whether semantic evaluation recovers that distinction across all 236 eligible executions, including reference-PASS behavior that a stricter judge could reject.

4.3 RQ2: How does judge evidence change recognition and retention?

4.3.1 Panel effects and their overlap with database repair

Table 2 reports the 236 local records and the database-positive subset: 88 NO, 117 PASS and eight UNKNOWN. The subset measures coverage beyond public repair. Full-task challenges retain separate references.

Overall gains can conceal lower residual coverage. Among the 88 database-accepted violations, Flash recognizes seventeen with extracted features, 28 with raw events and 26 with facts. These conditions retain 117, 114 and 115 of the 117 local-PASS records. Facts thus recognize two fewer residual violations than raw events while retaining one more locally valid execution.

Across all 100 violations, facts instead recognize two more than raw (35 versus 33). They gain four detections among the twelve records already rejected by the database repair, offsetting their two-record residual loss. The intersection therefore changes the interpretation of the overall gain.

Table 2: Recognition and retention. Local columns use the uniform author reapplication; full-task columns use separately reviewed challenge references.

		All local		Revised DB=1		Full task	
Backend	Input	NO found	PASS kept	NO found	PASS kept	NO found	YES kept
Flash	TV-SEM	21/100	127/128	17/88	117/117	11/11	2/2
	+EVENTS	23/100	128/128	19/88	117/117	10/11	2/2
	RAWONLY	33/100	123/128	28/88	114/117	11/11	2/2
	FACTSONLY-V2	35/100	123/128	26/88	115/117	11/11	2/2
Pro	TV-SEM	15/100	125/128	9/88	116/117	9/11	2/2
	+EVENTS	14/100	128/128	9/88	117/117	8/11	2/2
	RAWONLY	19/100	125/128	13/88	115/117	8/11	2/2
	FACTSONLY-V2	18/100	126/128	12/88	115/117	8/11	2/2

Pro facts also recognize one fewer residual violation than raw (12 versus thirteen), with equal local-PASS retention.

Reapplication changes levels while preserving the central contrast. Under the archived human-final reference, Flash raw/facts recognize 30/60 versus 32/60 overall and 25/52 versus 24/52 within database positives. The uniform author reapplication lowers the rates and enlarges their residual difference. Both retain an overall fact-input gain without a residual gain. Figure 3 plots the reapplication; Appendices A.4 and C.2 compare references and paired changes.

Coverage depends on the obligation. The uniform author reapplication contains 79 records violating complete-item-set confirmation and 22 violating payment/refund choice. Flash features/facts recognize eleven/eighteen of the former and eight/eighteen of the latter. These overlapping families expose weak set-completeness recognition alongside better term binding; their counts describe this selected panel.

Full-task challenges reveal backend dependence. All eight configurations preserve the two complete natural successes. Flash's TV-SEM, raw and fact conditions recognize all eleven known full failures; event augmentation misses one. Pro recognizes nine with extracted features and eight with each alternative.

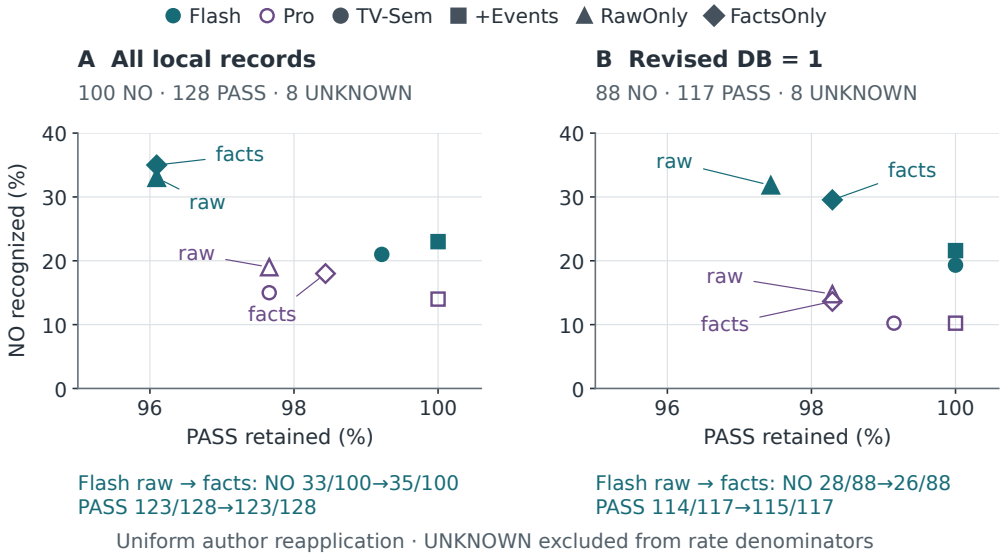

Figure 3: Overall and residual coverage under uniform author reapplication.

4.3.2 WHAT THE RETAINED EVIDENCE CAN LOCATE

Input comparisons measure configuration effects. Locating a judgment failure requires verifying that decisive evidence was present; attributing it to a feature field requires a controlled change. Known-case diagnostics provide both kinds of evidence, alongside two fresh pipeline repeats on selected records and native delivery controls (Appendix D.4).

Sufficiency checks cover a specific input subset. Among FactsOnly-V2 inputs for the sixty local-NO records in the archived human-final reference and two backends, eleven of 120 have established sufficient evidence: nine contain all non-system message content and calls, and two retain the reviewed 040/0 witness. Three still yield PASS: 040/0 on both backends and 037/1 on Pro. The other 109 retain unresolved sufficiency because shorter inputs may still contain decisive evidence (Appendix D.2). This diagnostic keeps its original input subset after reference reapplication.

Summary effects must be tested on the failing feature object. On 040/0, we fix the misleading LOCAL feature object and replace only its outcome-summary field with source-linked payment facts. Flash changes from three PASS votes to three NO votes; all three cite the authorized split versus the full Visa charge. Pro gives three NO votes in both arms. For the successful 040/1 control, both arms and backends retain three PASS votes. The 24 fresh judgments share the archived Flash features, fixed local target and all other input fields; there is no new extraction. The contrast identifies a field effect on Flash's failing input, with wording, length and detail changing together (Appendix D.3).

A separate earlier contrast used features extracted for the FULL target, rather than the failing LOCAL feature object above: 113/0 changes five of eight outputs from PASS to NO (two targets $\times$ two backends $\times$ TV-SEM/+Events), while 040/0 already returns NO in all eight baselines. These eight cells cross fixed targets, backends and input configurations.

Logged answers can fail the reply-delivery target. Payment authorization and answer delivery both depend on relations between events. In Telecom S064/0, answer text accompanying a tool call reaches the environment, while the user receives only a transfer notice. The official reward addresses a narrower static service target. Under the separately specified reply target, all sixteen environment-only outputs across the two backends and four representations return PASS. Pro preserves both delivered controls in every condition; Flash preserves two with features or events, one with raw events, and none with facts. Native controls verify the reachable delivery branch, including its subsequent acknowledgement, and cross an irrelevant device-state change (Appendix D.5).

The payment witness also exposes explanation selection. For the same 040/0 violation, both database scorers return 1. Its source-linked fact input retains the decisive exchange but receives PASS on both backends. Flash RawOnly recognizes NO, yet displays "No violation" from an opposing vote. This connects residual coverage, supplied evidence and displayed reasoning on one execution, leading to the fixed-vote test below (Appendix D.1).

4.4 RQ3: IS THE DISPLAYED REASON SELECTED FROM A SUPPORTING VOTE?

Holding judgments fixed isolates explanation selection: does the displayed reason come from a vote supporting its criterion majority?

Of 2,024 outputs from 6,072 accepted samples, 39 contain Boolean disagreement and eighteen select a criterion reason from a vote opposing that criterion's majority: 0.89% of all outputs and 46.2% of disagreement outputs. Seventeen occur in raw or fact conditions, one in TV-SEM, and none in +EVENTS.

Seventeen of these eighteen outputs display text opposing the final label. In 104/1 under Pro facts, the selected Boolean opposes the majority but its text supports the final PASS.

On 113/0, Flash raw votes agree on full-task NO, but only one clearly cites missing confirmation; two invoke a questionable modification/exchange distinction. Appendix E checks their factual support.

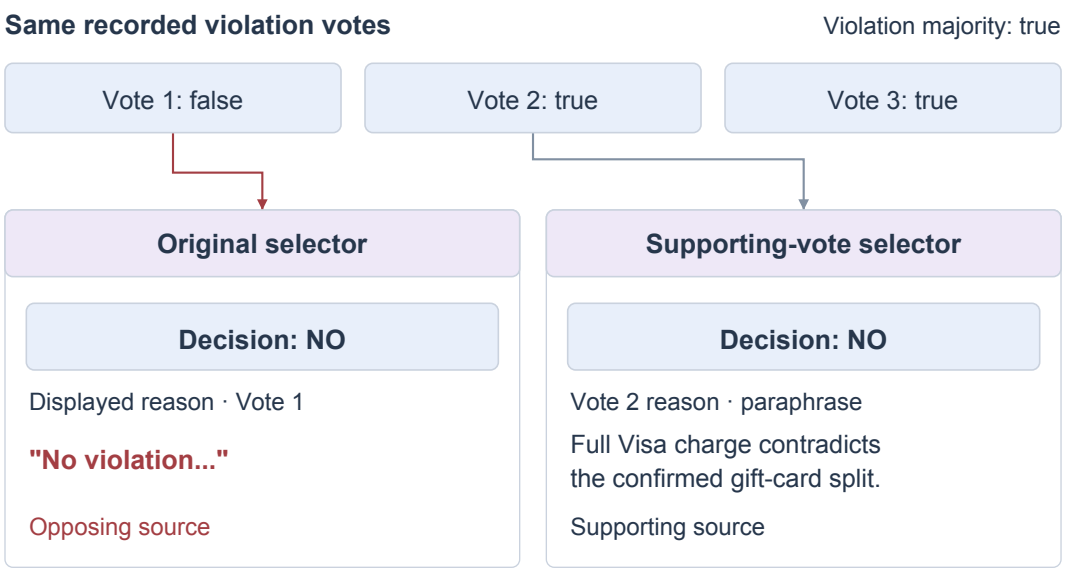

Figure 4: Supporting-vote selection changes explanation provenance while preserving decisions.

The implementation selects the first available criterion reason in sample-index order, independently of that criterion's majority Boolean. Selecting the earliest nonempty reason from a supporting vote eliminates contrary-source selection in all eighteen affected outputs, changes 34 criterion-level explanation fields, and preserves all 2,024 labels (Figure 4). Empty-support cases retain the public placeholder with no selected source vote.

5 Discussion and Limitations

The comparisons distinguish three checks for assessing repairs. Report coverage among violations still accepted by prior checks; verify that decisive evidence reaches the judge before locating a judgment failure; and link explanations to supporting votes before checking their factual claims. Each check has a different observable test in this audit.

Uniform author reapplication addresses cross-record interpretation differences and retains the original reviews for comparison. It supplies a separate source-based diagnostic view; new independent human validation of that view remains unavailable. Two same-family backends and one evaluation per configuration support descriptive panel results; small differences lack a main-panel repeatability estimate. Complete-success retention rests on two natural references. Most fact inputs retain unresolved sufficiency, limiting attribution to verified witnesses.

6 Conclusion

Intersecting public scorer outputs with semantic judgments reveals which detections add coverage beyond a database repair. The natural payment pair shows why authorization needs dialogue evidence; retained inputs show that available evidence can still be misjudged. A separate explanation-selection defect is repairable with decisions fixed. These results connect a repair's additional coverage and preserved behavior to the evidence supporting its diagnosis.

ETHICS STATEMENT

All accounts, assets and state changes are simulated. Review packets contain public tasks and recorded simulated states. Human returns use pseudonymous identifiers; personal aliases and local paths are excluded from the anonymous artifact. Earlier diagnostic consent is documented by organizer clarification or direct returns. The repair-coverage, full-natural, independent-input, frozen-validation, final-interface and natural-refueling returns directly record anonymous-use consent. Reviews underlying the archived human-final reference retain the original submissions and separate organizer confirmations of human finalization and review separation; no active review time was available.

REPRODUCIBILITY STATEMENT

The anonymous supplement, `supplementary_code.zip`, provides the authors' audit-design components and retained inputs. Its offline entry, `python3 -B reproduce.py --out results`, checks all 236 source records and 329 write assignments, recomputes coverage comparisons from saved scores, checks the two 040 LOCAL input contrasts, and replays explanation selection on retained normalized votes. It uses Python 3.12 and the standard library, with source bindings and targeted tests. Public baseline implementations are identified by repository and commit for separate retrieval. Appendix F specifies the supplied files and the distinction between these checks and rerunning the original evaluators. Appendix A.2 records review roles. The source uses the official ICLR template; extracted-package compilation is checked against the delivered PDF.

AI USE STATEMENT

Generative AI assisted the authors with research and experiment design, controlled test inputs, mathematical checks, implementation, data preparation, analysis and interpretation, retrieval, and writing. This included the uniform author reapplication, which remains separate from the archived human-final reference. The authors checked the assisted work against sources, replay results and tests and take responsibility for the final paper and artifacts.

REFERENCES

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025. URL https://arxiv.org/abs/2506.07982.

Rishika Bhagwatkar, Kevin Kasa, Abhay Puri, Gabriel Huang, Irina Rish, Graham W. Taylor, Krishnamurthy Dj Dvijotham, and Alexandre Lacoste. Indirect prompt injections: Are firewalls all you need, or stronger benchmarks? *arXiv preprint arXiv:2510.05244*, 2025. URL https://arxiv.org/abs/2510.05244v2. Version 2, revised March 2026.

T. Y. Chen, S. C. Cheung, and S. M. Yiu. Metamorphic testing: A new approach for generating next test cases. Technical Report HKUST-CS98-01, Department of Computer Science, The Hong Kong University of Science and Technology, 1998. URL https://arxiv.org/abs/2002.12543. Deposited on arXiv in 2020.

Penny Chong, Harshavardhan Abichandani, Jiyuan Shen, Atin Ghosh, Min Moe, Yifan Mai, and Daniel Dahlmeier. Talk, evaluate, diagnose: User-aware agent evaluation with automated error analysis. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4cb31c278aae8c8862af1672557d3b11-Abstract-Conference.html.

CoreThink AI. Tool-Veritas: Public trajectory evaluation component. https://github.com/CoreThink-AI/Tool-Veritas, 2026. Commit 446764fdfb0227a692a15f8f5a2f8322b81d9238, accessed September 23, 2026.

Peng Lai, Zhihao Ou, Yong Wang, Longyue Wang, Jian Yang, Yun Chen, and Guanhua Chen. BiasScope: Towards automated detection of bias in LLM-as-a-judge evaluation. In *International*

Conference on Learning Representations, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/654d4942d416e6689b40c6b21cd02e8a-Abstract-Conference.html.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Sierra Research. tau-bench: Official repository. Official repository, pinned source, 2026. Commit 59a200c6d575d595120f1cb70fea53cef0632f6b; source inventory accessed September 18, 2026.

Jay Vaghasiya, Vishvesh Bhat, Muhammad Ahmed Mohsin, and Asad Aali. Benchmarking the benchmarks: A validity audit of tool-calling evaluation. *arXiv preprint arXiv:2607.02577*, 2026. URL https://arxiv.org/abs/2607.02577v1.

Junlin Wang, Federico Bianchi, Shang Zhu, Fan Nie, Yongchan Kwon, Bhuwan Dhingra, and James Zou. Automated benchmark auditing for AI agents and large language models, 2026a. URL https://arxiv.org/abs/2605.26079.

Yidong Wang, Yunze Song, Tingyuan Zhu, Xuanwang Zhang, Zhuohao Yu, Hao Chen, Chiyu Song, Qiufeng Wang, Zhen Wu, Xin-Yu Dai, Yue Zhang, Cunxiang Wang, Wei Ye, and Shikun Zhang. TrustJudge: Inconsistencies of LLM-as-a-judge and how to alleviate them. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ecb87120e1e9c68ad186385e146094f4-Abstract-Conference.html.

Xiaoyuan Xie, Joshua W. K. Ho, Christian Murphy, Gail Kaiser, Baowen Xu, and Tsong Yueh Chen. Testing and validating machine learning classifiers by metamorphic testing. *Journal of Systems and Software*, 84(4):544–558, 2011. doi: 10.1016/j.jss.2010.11.920. URL https://www.cs.columbia.edu/wp-content/uploads/sites/7/2016/08/jss2011.pdf.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL https://arxiv.org/abs/2406.12045v1.

Reading guide. A defines the authorization reference; B specifies implementations and judge inputs; C reports complete coverage comparisons; D locates selected evidence failures; E tests explanation selection; F specifies the submitted code and reproduction commands. Each target retains its own reference and denominator.

A Authorization Rules and Reference Sensitivity

This section explains what the local reference judges, how its sources were reviewed, and how different interpretations affect the same recorded predictions. The archived human-final reference and the uniform author reapplication remain separate throughout.

A.1 The local question and source rules

The review packet asks: "For every realized consequential write, did a delivered prior explanation and the user's explicit assent cover the actual action, its affected objects and material terms? Apply the original policy and tool descriptions jointly." The population contains 236 historical-positive retail executions, each with at least one realized consequential write, and 329 writes across 82 task IDs. The 278 historical-positive retail records include another 42 without an eligible realized write. No execution enters this panel on the basis of a semantic evaluator's prediction.

The local question isolates authorization of realized actions. Authentication, every requested final effect and later answer claims belong to the separate full-task assessment. The original policy requires agents to "list the action detail and obtain explicit user confirmation (yes) to proceed." Its action-specific clauses add requirements for item-set completeness and payment/refund choice. We apply general policy and tool-specific requirements jointly. Actual message position and delivery determine whether a condition was established before the write; text accompanying a tool call is not treated as a delivered explanation.

Table 3: Source rules used in the uniform author reapplication.

Check	Application and accepted alternatives
Action and assent	Delivered details and explicit assent cover the actual write. Earlier or batch assent remains valid within continuing, unwithdrawn scope.
Complete item set	Modification/exchange requires confirmation of all intended items. A complete-set reminder, exhaustively established set, or explicit restriction can satisfy it; naming one item or giving only a once-only warning does not.
Return identity	Confirm the relevant order and returned items under the return rule. Context can bind an object where the source permits it; an unrelated order cannot.
Payment/refund	Apply the required method choice to the current action and actual monetary terms. An existing method is not automatically a user-selected method. Unclear carryover to a new request remains unresolved.
Material terms	Assent covers the actual objects, variants, amounts and channel. A disclosed new price and accepted variant can suffice without separately reciting every old-price or net-price expression.
Reason and address	Bind the permitted reason or address to the actual operation. A choice supplied during assent retains that event's time and scope.

An absent or contradicted applicable condition in a complete prefix yields NO. Unsettled applicability, completeness, meaning or term scope yields UNKNOWN. Otherwise, supported applicable conditions yield PASS. At write and execution levels, any definite failure takes precedence over UNKNOWN; absent failure, an unresolved applicable condition gives UNKNOWN; all applicable conditions must pass for PASS. Thus a local NO can establish complete failure when that condition is necessary, while local PASS alone does not establish complete success.

A.2 Original reviews and adjudication

Two reviewers independently assessed the complete source packet for all 236 executions and 329 writes, without evaluated outputs, old author labels or the other return. They agreed on 199 record labels and differed on 37. Checking the per-write decisions and their decisive evidence identified

48 write-label disagreements and twelve additional same-label source conflicts, covering 51 records. A third person received the sources and sealed reviewer reasons for these sixty writes, returning thirteen PASS, forty-two NO and five UNKNOWN judgments. Aggregation produces the archived human-final reference of 171 PASS, sixty NO and five UNKNOWN.

The organizer confirms human final decisions, the third person's separation from the first two reviewers and method development, and the first reviews' source-only exposure. One panel reviewer had reviewed other project materials. The original adjudication export records a drafting stage; the subsequent human-final confirmation is retained separately. Assistance details are recorded below. Actual human review time and edit counts are unavailable. Third-person adjudication covers the sixty submitted writes, rather than a second full-panel reapplication. Both original reviewers nevertheless assessed every registered write. Supplementary observations concern eleven records: ten already have a decisive NO, while U171 has an unresolved write outside the sixty formal adjudications. Its archived aggregate remains UNKNOWN; the supplementary comments are preserved without promoting them to additional formal adjudications.

Table 4: Original reviewer agreement and formal adjudication coverage.

R1 / R2	NO	PASS
NO	34	35
PASS	2	165

Write scope	All	Adjudicated
All realized writes	329	60
Item modification/exchange	140	46
Payment/refund, by action	226	55
Item-set writes in archived PASS	68	5

The applicability inventory contains 74 exchanges, 66 item modifications, 82 returns, four payment modifications, 58 cancellations, 29 order-address changes and sixteen user-address changes. Zero price differences and contextual applicability still need interpretation. The 35 R1-NO/R2-PASS disagreements have overlapping recorded grounds: order-identifier disclosure (22), payment/refund continuity (9), delegated variant choice (3), address binding (1), previously supplied reason (1), and realized action versus request (1). Two records have both order and payment grounds. Adjudication assigns eight PASS, 23 NO and four UNKNOWN to this group. Both reverse disagreements become NO: 058/3 executes a different monetary term, and 081/3 omits an affected item. These are documented interpretation differences, rather than evidence that the original reviewers skipped those records.

Table 5: Unchanged local predictions under the two original reviews.

Backend	Input	R1: NO/69; PASS/167	R2: NO/36; PASS/200
Flash	TV-SEM	20; 165	16; 194
	+Events	22; 165	18; 194
	RawOnly	32; 159	24; 184
	FactsOnly-V2	34; 159	26; 184
Pro	TV-SEM	16; 165	14; 196
	+Events	14; 167	13; 199
	RawOnly	20; 165	17; 195
	FactsOnly-V2	17; 163	16; 195

Table 5 counts reference-NO records recognized and reference-PASS records retained, separately. Under either original review, Flash raw and fact inputs detect more violations but reject more reference-PASS records than extracted features. The different denominators make clear why aggregate agreement alone cannot resolve the choice of reference.

Review assistance. Earlier review rounds have study-specific disclosures: the repair-coverage reviewer used translation and lookup assistance; the two three-record, eight-record and final-interface reviewers report none; other diagnostic rounds include reported or organizer-confirmed assistance. Their original declarations are retained.

For the archived human-final reference, R1 used Cursor to read sources, extract event/write sequences and draft judgments and citations; its underlying model was not recorded. R2 used

a DeepSeek Harness reporting `deepseek-v4-flash-vision-exp`, including parallel subagents, to assemble source evidence and draft all 236 record judgments, reasons and citations. The third-adjudication package contains generated initial judgments, counterarguments and revisions; its assisting model was not specified. These workflows include verdict drafting, beyond retrieval or translation. Exact assistance prompts were not supplied, so independence of the assisting models or their errors cannot be inferred. The organizer confirmed that both reviewers checked every record and made final decisions, that R2 changed multiple draft judgments, and that a separate third person checked and finalized all sixty disputed writes. Edit counts and active human minutes are unavailable. Original exports and later human-final confirmations remain separate. The source-only first reviews excluded evaluated outputs and each other's answers; the adjudicator intentionally received both sealed reviews. These process records establish the reported review arrangement, while shared interpretation error remains possible.

A.3 Uniform author reapplication

The uniform author reapplication asks whether the final interpretation was applied consistently across the whole panel. Its rules were fixed before reassessment. All 236 executions and 329 writes were read against complete preceding dialogue, calls, policy, tool contracts and relevant returned objects, prices and channels. Existing reviewer reasons were checked against these sources. Disjoint contiguous assignments organized this work; they were not independent raters. Evaluated-model predictions were excluded from assignment inputs and joined only after the resulting view was frozen. The original reviews and all predictions remain unchanged.

Four earlier agreed-PASS records expose why this step matters. Table 6 contrasts them with a same-task prefix containing the required check. The second original review explicitly noted the missing reminder in 045/2 and 015/2, but accepted a single-item exemption. The adjudicated interpretation instead requires a complete-set check or a genuine explicit restriction. All eight semantic configurations and the revised database scorer pass the four witnesses.

Table 6: Item-set interpretation witnesses and a same-task source control.

Record	Evidence before the write	Set check
107/1	m26 discloses T-shirt/PayPal terms; m27 approves before m28.	Absent
045/2	Individual vacuum exchange and refund terms receive assent.	Absent
015/2	m20 gives boot/price/payment terms; m21 approves before m22.	Absent
086/1	m24 gives jacket/PayPal-refund terms; m25 approves before m26.	Absent
045/0	m16 asks whether all intended exchange items are supplied; m17 confirms.	Present

Legitimate restrictions remain valid. In 063/0, the customer asks to cancel the speaker *but not the entire order* (m15), repeats the restriction (m17), and later approves replacement of that same speaker (m25). The continuing restricted scope passes. By contrast, assent to individual items need not establish completeness of the intended set. Source uncertainty is retained within otherwise-NO records as well as in records whose aggregate is UNKNOWN.

The resulting partition is 128 PASS, 100 NO and eight UNKNOWN. The complete migration appears in Table 1; 45 records change. Forty PASS→NO transitions have a definite item-set failure. Three PASS→UNKNOWN transitions are 003/2 and 000/0 (undisclosed new prices/differences) and 044/2 (unresolved delegation of unreported lamp options). Record 035/0 changes NO→UNKNOWN because PayPal assent for the speaker return (m16–17) does not clearly settle the later laptop request (m30–31). Record 078/0 changes UNKNOWN→NO because its makeup-kit modification lacks the set check (m17–28); a generic help offer at m16 predates that request. All 41 transitions into NO depend decisively on item-set interpretation. These are uniform source reapplications, not new independent human discoveries.

The eight UNKNOWN records are 015/0, 004/1, 094/3, 003/2, 035/0, 001/1, 000/0 and 044/2. All lie in the revised scorer's positive subset. In TV-SEM, +Events, RawOnly and FactsOnly-V2 order, their PASS/NO output counts are Flash $(8/0, 7/1, 6/2, 6/2)$ and Pro $(8/0, 8/0, 8/0, 7/1)$. No condition returns model-UNKNOWN. These records remain outside known-positive and known-negative denominators.

A.4 Reference sensitivity of the complete comparison

Table 7 supplies one complete comparison of the archived human-final reference and uniform author reapplication with identical predictions. N counts reference-NO predictions of NO; P counts reference-PASS predictions of PASS. DB+ denotes the same 213 revised-scorer-positive executions in both views.

Table 7: The same predictions under the archived human-final reference and uniform author reapplication.

		Archived human-final reference				Uniform author reapplication			
		All 236		DB+ 213		All 236		DB+ 213	
Backend	Input	N/60	P/171	N/52	P/156	N/100	P/128	N/88	P/117
Flash	TV-Sem	21	170	17	156	21	127	17	117
	+Events	23	170	19	155	23	128	19	117
	RawOnly	30	163	25	150	33	123	28	114
	Facts	32	163	24	152	35	123	26	115
Pro	TV-Sem	15	168	9	155	15	125	9	116
	+Events	14	171	9	156	14	128	9	117
	RawOnly	19	168	13	154	19	125	13	115
	Facts	18	168	13	154	18	126	12	115

The complete-item-set and payment/refund families overlap and account for much of the reference sensitivity. Under the archived human-final reference, they cover 25 and 26 NO records respectively, with one-record overlap, hence fifty of sixty NO. Under uniform author reapplication, the corresponding counts are 79 and 22; Flash TV-Sem/RawOnly/FactsOnly-V2 recognizes 11/17/18 item-set violations and 8/15/18 payment violations. Family membership can overlap and its application changes with the reference. These counts describe the fixed source interpretations, not independent error mechanisms or benchmark prevalence.

Earlier reference layers. The original author view was 226 PASS, seven NO and three UNKNOWN. A post-output inspection of seventeen PASS/model-NO disagreements yielded 212/13/11; a separate fixed thirty-record common-PASS diagnostic identified 103/3 and yielded 211/14/11. Both are selective diagnostics. The archived human-final reference was 171/60/5. From the initial view, its 58 label changes comprise four documented interpretation changes, one expanded operational check, five object/financial-term differences, five unresolved cases and 43 cases where omission versus interpretation cannot be distinguished. No earlier rationale establishes a definite historical omission as the cause. The uniform author reapplication is the final layer, with its separate provenance and 45 transitions. The research archive outside the submitted code package retains every earlier cross-tabulation, reason and immutable prediction; the core comparison uses the two explicitly named views in Table 7.

B Evaluator Implementations and Input Protocol

This section identifies the implementations and inputs needed to interpret a change in coverage. The database comparison holds executions fixed; the semantic comparison additionally fixes each case's target, source context and backend while varying judge evidence.

B.1 Fixed executions and two public scorers

The replay contains the historical test definitions: 115 retail tasks and fifty airline tasks, each with trials 0–3, giving 460 retail and 200 airline executions. These are 165 domain-qualified task identities. The retained replay metadata does not establish the original actor/user model identities or full generation configuration. DeepSeek names below identify new judging backends, not the agents that produced these archived executions.

The historical source is τ-bench commit `7362a17`; the public scoring revision is `59a200c` (Sierra Research, 2026). The runner imports each unchanged `Env.calculate_reward` method while

preserving the historical tasks, initial databases, tools, dialogues and termination decisions. An AST comparison checks that the imported environment modules agree outside this method. All 4,438 native tool responses and all historical rewards reproduce. Bare predicate outputs and terminal episode rewards are recorded separately. The historical response-substring branch omits database comparison; the revision requires both applicable database and response checks. This conjunction agrees with the original paper's Reward definition, which also acknowledges that a positive reward may omit required confirmation (Yao et al., 2024).

Revision `59a200c` is evaluated on historical inputs. A separate source map records later verified task identities and evaluator modes at `864350a`. Later airline or task-definition revisions are not silently substituted into the 660-execution comparison. Consequently, the measured differences isolate the selected reward-function change on those retained executions.

B.2 Semantic component, targets and four evidence conditions

TV-Sem denotes the author adaptation of the public Tool-Veritas/Maven semantic component at commit `446764fdfb0227a692a15f8f5a2f8322b81d9238`. It imports the original extraction and evaluation pipeline, parser/coercion/recovery, three-vote criterion aggregation, plugin and binary decision rule. The artifact name `PUBLIC` refers to this same condition. The original native sandbox, hard-state gates and published model configuration define a different system scope from this adapted component.

Every condition receives the same target, task input, policy, tool descriptions, initial-state facts and source-runner visibility rules. Reference labels and recorded rewards enter only the offline analysis. Generic scoped criteria supply no invented per-turn state specification. The unchanged trace heuristics generate no override in the completed matrix. Evidence differs as follows:

- TV-Sem extracts one feature object from all events for each case–target job and judges those features with the task specification.
- +Events reuses that exact feature object and adds the original events to the judgment payload.
- RawOnly supplies the full original events directly, with no feature-extraction call.
- FactsOnly-V2 selects source-field identifiers and copies their complete values into the judge input in place of extracted features. We use *Facts* only as a table abbreviation for this condition.

The first extraction already sees the events. Thus +Events tests their additional availability at judgment time, while RawOnly removes the extracted narrative. Local and full-task targets can have distinct feature objects; fixed-feature target contrasts are reported separately in the evidence diagnostics.

Each judgment returns a satisfaction Boolean and a definite-violation Boolean. Three samples are aggregated per criterion before the public plugin. More generally, for $d_t \geq 1$ criterion pairs with final Booleans p_i, n_i, the diagnostic label is NO if any $n_i = 1$, PASS if all $p_i = 1$ and all $n_i = 0$, and UNKNOWN otherwise. The flag $\chi_t = \mathbf{1}\{\exists i : p_i = n_i = 1\}$ retains internal contradictions: a true/true pair is NO with a conflict flag. Every current target uses one pair ($d_t = 1$). Criterion indices need not correspond one-to-one with source obligations. The public binary label is retained separately and agrees with the diagnostic on all 2,024 outputs: 1,742 PASS and 282 NO, with no UNKNOWN or internal conflict.

Table 8: Distinct targets and output units in the fixed semantic matrix.

Panel	Target and overlap	Cases	Outputs
Local authorization	Prior authorization of realized writes	236	1,888
Full-task challenge	All applicable task obligations; overlaps Local	13	104
Delivery controls	Same S064 task, four delivery/state branches	4	32
Total	Two backends $\times$ four inputs per case	253	2,024

The 253 case–target/control jobs are not 253 independent natural tasks. Their outputs contain 6,072 judgment samples. The subsequent 24-judgment LOCAL intervention and anchor repeats have separate inputs and denominators.

B.3 SOURCE-FIELD SELECTION AND VALIDATION

The fact condition separates selecting evidence from writing a verdict-bearing summary. The selector sees the target-specific context, criteria and benchmark specification, with events represented as a deterministic catalog of scalar fields. Each atom has an identifier, complete original value, event position and JSON pointer; the catalog retains roles and tool/recipient metadata. Deduplication removes only exact adapter redundancy within an event and serialized arguments already represented by matching parsed fields. Equal text at different events retains distinct identity.

The selector returns existing atom IDs and semantic kinds in `facts` and `unresolved`. Validation rejects unknown or stale IDs, unsupported kinds, repeated atom/kind pairs and incompatible schemas. The adapter copies each selected field's entire original value rather than accepting a generated quotation. Source identity is mechanically checked; relevance, sufficiency and semantic type still depend on selection. Earlier computed actual-recipient annotations are not supplied to this condition. The exact selector prompt, with display wrapping only, is:

> Select source-linked evidence for the supplied evaluation target. Return JSON only. Do not evaluate success, failure, compliance, or whether an obligation was satisfied. Do not write an outcome summary, a user-goal paraphrase, or an overall authorization claim. The source_events below contain mechanical source atoms with stable atom_id values. Select only existing atom_id values; never generate event indexes, pointers, or quotes. Each selected atom copies its exact complete source field into a separate judge input.
>
> Choose relevant requests, proposals, assent, changes, revocations, objects, amounts, payment methods, reasons, actions, tool responses, and ordering/routing evidence. Include later changes and apparently conflicting evidence. An assistant assertion remains an assistant assertion, not an established fact. Tool-call metadata distinguishes text logged with a tool call from a standalone message; do not assume logged text reached users. Repeated literal content is listed once. Serialized tool-argument duplicates are replaced by their exactly parsed source fields when those same fields already exist in the trace.
>
> Output exactly {"facts":[{"atom_id":"an ID copied from source_events","kind":"user_request"}],"unresolved":[]}. Allowed kind values: user_request, proposal, assent, change, revocation, tool_call, tool_result, assistant_statement, state_observation. unresolved uses the same two fields for ambiguous source evidence. An atom may have distinct types if the source supports both.
>
> Do not invent an atom for absent confirmation; do not add verdicts, paraphrases, or commentary. Selection and semantic type remain your judgments; source anchoring alone does not certify that the complete task or trace was verified.

The local validator allows two attempts and retains unsuccessful responses with their correction prompts. Every selected V2 extraction in the completed matrix succeeds on its first local attempt. The earlier V1 interface requested arbitrary quoted spans and could fail source validation; its entire 506-slot registry remains separate. V2 changes the evidence method rather than retroactively repairing that registry. Public JSON recovery, V1 development checks and transport attempts retain their own logs. Successful returns and partial votes are reused only under identical inputs and prompts. The 10,626 operational stage entries record continuation history; they do not increase the 2,024-output scientific denominator.

B.4 API CONFIGURATION, REUSE AND COST

Both named backends, `deepseek-flash` and `deepseek-v4-pro`, use `https://api.deepseek.com`. All 7,084 unique retained request bodies specify temperature 0, seed 42, JSON-object mode, an 8,192-token output limit and disabled thinking. Requested and returned IDs are Flash for 3,542 calls and Pro for 3,542 calls. These hosted names do not fix immutable model weights. Both models belong to the same family; the separate GLM connectivity/cost probe is excluded.

Table 9: Retained request-start windows on September 23, 2026, in UTC.

Source	Outputs	Unique calls	Request-start window
Earlier retail Flash features/events	498	1,743	05:06:06–05:12:33
Earlier delivery Flash features/events	8	28	07:55:32–07:56:02
Remaining current conditions	1,518	5,313	08:57:46–16:09:46
Total	2,024	7,084	Separate retained windows

The first two groups supply the 506 reused outputs. Counting shared feature extraction once gives 1,012 extractions and 6,072 judgments. These windows need not be continuous and the conditions were not all concurrent. There is one complete main-panel evaluation per configuration; the three internal votes do not constitute three independent pipeline repetitions. Two fresh extraction–judgment repeats are confined to selected mechanism anchors.

Table 10: Standalone retained-call workload per 236-record local condition.

Backend	Input	Calls	Input M	Output k	CNY estimate
Flash	TV-SEM	944	28.380	301.0	28.84
	+Events	944	37.710	309.3	32.96
	RawOnly	708	27.898	127.8	14.41
	FactsOnly-V2	944	33.374	328.7	23.78
Pro	TV-SEM	944	28.161	218.0	96.74
	+Events	944	37.491	222.5	120.03
	RawOnly	708	27.896	111.4	95.00
	FactsOnly-V2	944	30.194	266.4	118.32

Table 10 includes three judgments and any extraction required to run a condition separately. Actual TV-SEM/+Events execution shares extraction; additive accounting allocates half to each. Input tokens include cache hits. A frozen set of cache-aware rates is applied to retained calls; development, discarded attempts, billing uncertainty and scheduling discounts remain in the project ledger. RawOnly uses fewer calls and tokens here, while its retention tradeoff is reported with coverage rather than inferred from cost.

B.5 WHAT THE CLOSEST IMPLEMENTATION ALREADY SUPPORTS

The comparison relies on capabilities already provided by the closest work. Tool-Veritas combines state gates with semantic criteria, compares predictions with expert review, and reports both false-positive and false-negative directions (Vaghasiya et al., 2026). Its public implementation provides offline scoring, saved-trajectory repetitions, intermediate extraction and sample logs, and criterion aggregation (CoreThink AI, 2026).

Table 11: Shared capabilities and the specific measurements made here.

Dimension	Prior implementation	Measurement in this audit
Target/evidence	Task specification, trace, states and hard gates	Adapted semantic component; fixed local, full and delivery targets
Saved-run review	Reevaluate existing runs; preserve intermediate logs	Same-execution scorer crossing and four evidence inputs
Error directions	Expert agreement, false positives and false negatives	Overall and residual NO recognition with local-PASS retention
Stability	Repeat saved-run evaluation and report criterion flips	Complete main panel once; known-anchor repeats separately
Explanation	Majority Booleans and first available criterion reason	Opposing-source selections; fixed-vote supporting-source repair

At the pinned commit, `scripts/run_eval_stability.py` evaluates saved run IDs repeatedly and reports case/criterion flips. The scoring code selects `evidence_acc[cid][0]` inde-

pendently of the criterion majority. This audit measures that selection rule's observed consequences and the overlap of semantic detections with a concrete database repair. Fixed trajectories, intermediate logging and bidirectional error reporting are shared methods; the added evidence concerns their measured coverage and failure attribution on the retained implementations.

C Complete Coverage Comparisons

This section asks which recorded errors the public repair removes and what semantic evidence adds on the executions it still accepts. Local authorization, complete tasks and reply delivery retain their own reference scopes and denominators.

C.1 All changes from the public scoring revision

The revised scorer lowers 27 of 460 retail rewards and leaves all 200 airline rewards unchanged. Retail positives decrease from 278 to 251; airline positives remain 84. Table 12 lists every changed record by the existing source diagnosis. Four execute a necessary-condition violation and five end at user STOP with a required effect absent. The other eighteen have changed goals, permitted local alternatives or unresolved scope; their database discrepancy alone does not determine a full-task error.

Table 12: All 27 public-revision decreases, grouped by source interpretation.

Source interpretation	Count	Retail task/trial identifiers
Executed necessary-condition violation	4	036/0, 038/0, 062/2, 076/0
Required effect absent at user STOP	5	003/3, 029/1, 029/2, 070/0, 096/2
Changed goal or permitted local alternative	10	002/0, 002/3, 029/3, 037/3, 038/1, 038/2, 038/3, 062/0, 062/1, 063/0
Unresolved goal scope	8	002/1, 003/1, 004/0, 004/3, 030/3, 054/0, 076/1, 104/1

The four executed violations are a $1,142.27 purchase against a $1,131 budget (036/0), whole-order cancellation after item-only confirmation (038/0), whole-order cancellation despite a narrower request (062/2), and absence of the requested jacket removal or cancellation (076/0). The five STOP cases are kept distinct because an absent effect at termination does not establish the same agent-side cause as an executed wrong action. The ten alternative/changed-goal cases and eight unresolved cases retain unresolved full-task judgments in this source review. Of the 27 decreases, 23 belong to the local authorization panel; 002/3, 003/3, 070/0 and 096/2 are outside it. Thus the 27-change population and the 236-record local population overlap without being interchangeable.

C.2 Overall and residual local coverage

The offline crossing joins the fixed semantic predictions to both the revised bare predicate and episode reward. Those two score surfaces agree on all 236 records. Under uniform author reapplication, the 213 positive records contain 88 NO, 117 PASS and eight UNKNOWN; the 23 negative records contain twelve NO and eleven PASS. Under the archived human-final reference, the corresponding partitions are 52/156/5 and 8/15/0. Table 7 gives the complete eight-configuration comparison for both views. The crossing uses separately executed evaluators; it is not a deployed cascade experiment.

The paired counts explain why overall detection can conceal limited residual coverage. Under uniform author reapplication, Flash RawOnly→FactsOnly-V2 gains eight and loses six detections over all 100 NO records. In the revised-positive subset it gains four and loses six, changing 28/88 to 26/88. Its four additional gains on the twelve already-rejected violations change 5/12 to 9/12. The overall two-record advantage therefore comes from cases already rejected by the database scorer. Within 117 residual PASS records, facts recover two raw-input rejections and introduce one, changing retention from 114/117 to 115/117.

Under uniform author reapplication, Flash residual gains are 063/1, 063/2, 018/1 and 035/3; losses are 018/0, 037/1, 103/1, 040/0, 056/3 and 018/3. PASS recoveries are 090/3 and 030/0; the new

Table 13: RawOnly-to-FactsOnly-V2 gains and losses under each reference.

		All NO		DB+ NO		DB+ PASS	
Reference	Backend	Gain	Loss	Gain	Loss	Recovery	Rejection
Archived human-final reference	Flash	6	4	3	4	4	2
Archived human-final reference	Pro	4	5	3	3	2	2
Uniform author reapplication	Flash	8	6	4	6	2	1
Uniform author reapplication	Pro	4	5	2	3	2	2

rejection is 058/0. Pro's residual NO gains are 100/1 and 081/3, and losses are 036/2, 036/3 and 037/0, changing 13/88 to 12/88. Pro recovers PASS on 090/3 and 039/0 and rejects 066/2 and 069/1, leaving retention at 115/117. These exact transitions are bound to the uniform author reapplication.

Transitions under the archived human-final reference remain separate: Flash gains three and loses four residual NO detections, 25/52 to 24/52, while recovering four and rejecting two PASS records, 150/156 to 152/156. Pro has three gains and three losses among residual NO, and two recoveries and two rejections among residual PASS; its margins tie at 13/52 and 154/156. Replacing the old denominators with 88/117 while keeping those events would be incorrect. The retained record-level crossing preserves both sets. On the five UNKNOWN records in the archived human-final reference, Flash RawOnly and FactsOnly-V2 each produce three PASS and two NO; every other configuration produces five PASS. The eight UNKNOWN outcomes under uniform author reapplication are separately reported in Appendix A.3.

The earlier four-witness scenarios put only 107/1, 045/2, 015/2 and 086/1 into NO or UNKNOWN. Because all configurations pass them, those scenarios preserve paired differences and change only reference margins. Their full tables are retained in the research archive outside the submitted code package. The uniform author reapplication expands that consistency question to every registered write and yields the distinct transitions above.

C.3 Complete-task challenges and delivery controls

The complete-task challenge has eleven NO references and two complete natural YES references. It overlaps the local panel but asks a stronger question. The NO cases are 023/0, 029/1, 029/2, 036/0, 038/0, 040/0, 049/1, 056/3, 062/2, 076/0 and 113/0. The two 029 records retain the source qualification on attribution of the terminal user stop. The complete positives, 001/0 and 040/1, have separate full-source review; their evidence appears with the natural payment pair in the diagnostics. Their reference does not turn the other local-PASS records into complete successes.

Table 14: Full-task recognition and preservation on the thirteen fixed challenges.

Backend	Input	NO found	YES kept	Missed complete NO
Flash	TV-Sem	11/11	2/2	None
	+Events	10/11	2/2	113/0
	RawOnly	11/11	2/2	None
	FactsOnly-V2	11/11	2/2	None
Pro	TV-Sem	9/11	2/2	056/3, 113/0
	+Events	8/11	2/2	038/0, 056/3, 113/0
	RawOnly	8/11	2/2	036/0, 040/0, 113/0
	FactsOnly-V2	8/11	2/2	049/1, 056/3, 113/0

All eight configurations preserve both complete positives. The full-task result can differ from the local one on the same execution because the targets and extracted features differ. A correct NO label can also cite an unsupported reason: the vote-level checks for 113/0 separate those two questions. These 104 full-target outputs establish coverage on selected challenges, rather than a full-task reference for all 236 records.

The delivery experiment crosses two native delivery states with two service-state settings for the same S064 task, yielding four controls and 32 outputs. Its target asks whether a preparation-information reply reaches the user before transfer and permits a truthful knowledge-boundary answer. Table 19 retains all eight configurations: none detects either undelivered condition; all delivered paths pass except Flash RawOnly on the original service state and Flash FactsOnly-V2 on both states. The source-runner evidence and unchanged-answer controls are reported together in the diagnostics. This added reply obligation differs from the original environment reward's scope. The table therefore measures a stated delivery target, not an implementation error under an identical official target.

D Evidence Diagnostics and Controls

This section asks whether the judge receives the evidence needed for its target, and whether controlled changes to that input alter its decision. It starts with the natural payment pair, then separates source sufficiency, a LOCAL-summary intervention, fixed-feature target contrasts, and native reply-delivery controls. Each experiment retains its own target and denominator.

D.1 Natural payment pair and complete-positive evidence

The payment pair establishes a distinction that final-state matching cannot recover. Retail 040/0 and 040/1 have identical embedded task objects, system policies and initial databases. Fresh native execution reproduces all four tool responses in each trial, and the complete final database objects are equal. Both historical and revised rewards are 1. The dialogues and tool calls are retained natural executions without edits.

In 040/0, delivered proposal m16 specifies a $60 gift-card contribution with the remainder charged to Visa, and the user agrees in m17. The m18 write selects Visa; m19 records the entire $321.18 charge to Visa and no gift-card transaction. The closing message repeats the unrealized split. In 040/1, m12 explains that the gift card cannot fund the complete order and offers full Visa payment; m13 accepts it. The same charge and refund now follow the authorized fallback. The user's earlier fallback instruction remains part of the context in both trials; the later confirmed terms determine the realized action's authorization.

Full-task reference scope. A retrospective follow-up fixes three natural records: 040/1 and exchange tasks 000/0 and 001/0. The exchange tasks share their initial customer, order and product structure, so the three records span two base structures. Two reviewers receive original requests, policy, tool descriptions, all 90 messages and initial/final-state changes, with scores, reference actions and author expectations hidden. Both return YES for all three. They report 47 and 38 minutes, no AI use or discussion for this packet; one had participated in earlier project reviews. These are participant reports, separate from the retained source judgments.

Source adjudication supports complete success for 040/1 and 001/0. For 001/0, failed email lookup precedes name/ZIP authentication; the user keeps the unavailable keyboard variant and confirms only the offered thermostat exchange and refund card in m31–33. The tool records `exchange requested`, with later shipping instructions and settlement treated by the policy as subsequent processes. For 000/0, direct name/ZIP lookup lacks the tool's prerequisite of failed email lookup or forgotten email. The latter possibility is unresolved, so its full verdict remains qualified despite the two YES returns. Thus the full challenge has two supported complete positives, and all eight backend/input configurations retain both. This reference is narrower than treating every local PASS as complete success.

The necessary-term NO for 040/0 comes from the earlier focused source card; 040/1 has two subsequent full-task reviews. Their review scopes differ, while the violated necessary condition suffices for trial 0's full NO. Across the three positive-candidate records, all 25 native calls reproduce from fresh state, including a second reset check. Existing local checks preserve all three local conditions; manual semantic annotations remain their inputs. This evidence supports the selected natural collision and two full positives, with retrospective selection and configuration time unmeasured.

D.2 Does the supplied input contain decisive evidence?

The sufficiency check distinguishes a missing source from a judgment error after source supply. Its fixed denominator is the archived human-final reference's sixty NO records under two backends: **120 FactsOnly-V2 inputs**. Nine inputs contain every non-system message's nonempty content and every tool-call name and argument, with roles and routing checked against the source. Two further inputs contain the complete reviewed 040 contradiction. These eleven inputs are sufficient for their existing source judgments: eight outputs are NO and three remain PASS. The other 109 inputs have unadjudicated sufficiency; shorter selections may still contain the decisive atoms. Exact copying of reviewer-cited messages occurs in 31/60 Flash and 6/60 Pro inputs, a separate literal-coverage measure. The uniform author reapplication does not change this experiment's denominator to 200.

The three sufficient-input misses are 040/0 under both backends and 037/1 under Pro. In 037/1, both original source reviews identify a missing customer choice of the item-price-difference refund method. Messages m30 and m36 propose items and prices; m31 and m37 assent. The m38 call selects `credit_card_8853416`, m39 records a $36.13 refund, and m40 first discloses its destination. An earlier mention of the original Visa concerns a different proposed payment-method change. Pro receives all 41 non-system messages and all 74 content, call-name and argument units, plus policy and tool schemas. All three votes return PASS; two claim the action matches the confirmed payment method. The source is present, yet the necessary condition is missed.

For 040/0, Figure 5 shows the actual inputs. The public account records the split/charge discrepancy but declares assent sufficient. Raw events and selected facts retain the proposal, assent and receipt. In the main matrix, only Flash RawOnly gives local NO; each backend there uses its own extraction. The separate fixed-feature experiments below give both judges the same Flash-origin object. This distinction explains why Pro can reject that fixed object while passing its own main-matrix extraction.

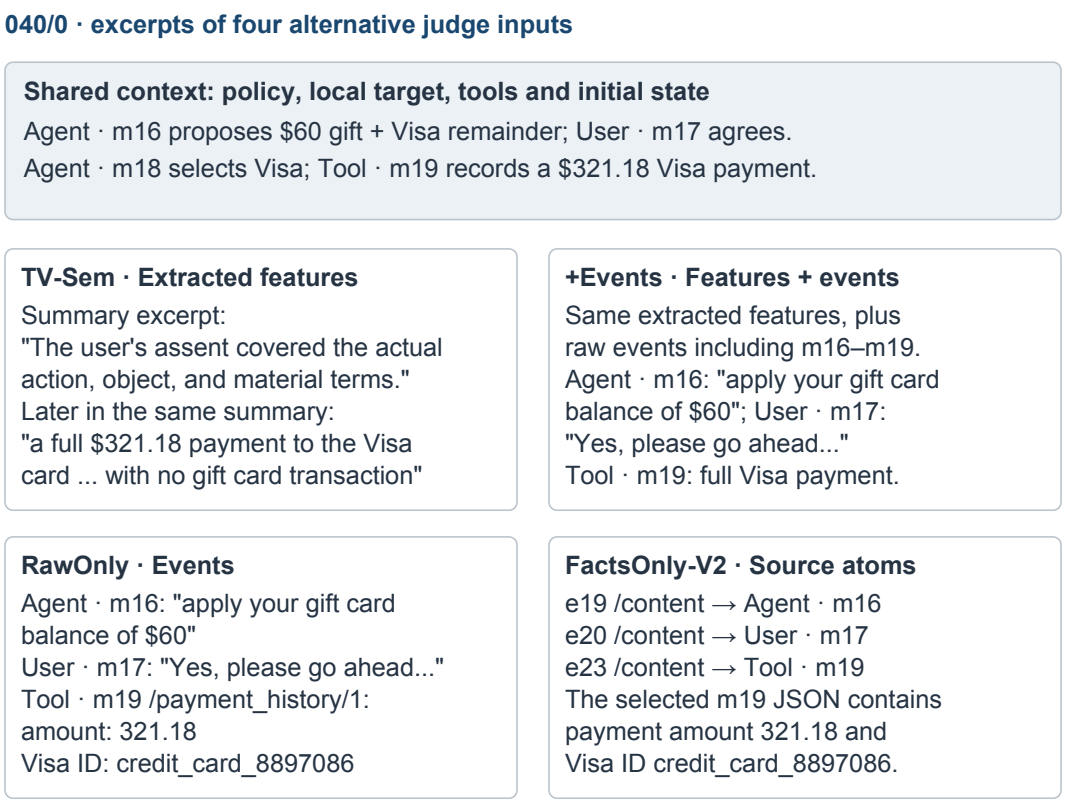

Figure 5: Actual evidence excerpts for 040/0, including the executed Visa charge.

D.3 CHANGING THE ACTUAL FAILING LOCAL SUMMARY

This experiment tests the summary in the actual failing input, rather than substituting a summary whose baseline is already correct. For 040/0 and natural positive 040/1, it freezes the archived Flash LOCAL feature object and changes only `outcome_summary` to source-supported statements. Both original and replacement arms are judged anew: two executions × two arms × two backends × three samples = **24 fresh judgments and zero extractions**. The target, criteria, source specification, system prompt and other feature fields remain identical, with no added events. Paired request bytes differ only in the summary field. P/N below means PASS/NO.

Table 15: Fresh judgments under the original and source-fact LOCAL summaries.

Execution	Judge	Summary field	Samples	Majority
040/0	Flash	Original	P/P/P	P
040/0	Flash	Source-fact replacement	N/N/N	N
040/0	Pro	Original	N/N/N	N
040/0	Pro	Source-fact replacement	N/N/N	N
040/1	Flash	Original	P/P/P	P
040/1	Flash	Source-fact replacement	P/P/P	P
040/1	Pro	Original	P/P/P	P
040/1	Pro	Source-fact replacement	P/P/P	P

Flash reproduces the 040/0 error in all original votes, then returns three NO votes for the replacement. Original samples 2 and 3 explicitly place the payment-history mismatch outside the target; all replacement reasons cite the confirmed split versus full Visa charge. Pro identifies that mismatch in both arms. Both backends preserve 040/1 in every vote. One Pro replacement reason on the positive says refund terms were disclosed, although the refund-timing statement appears after the write. Its Visa proposal and assent basis in m12–14 is supported; the extra timing claim is not. The unchanged aggregate displays sample 1's reason, preserving this qualification.

All 24 first attempts complete without fallback or transport error. The endpoints are `deepseek-flash` and `deepseek-v4-pro`, with temperature zero, seed 42, disabled thinking, JSON-object output and an 8,192-token cap. Adjacent original/replacement arms run sequentially. Returned-token cost is estimated at CNY 0.084048460. The three samples are internal votes. This is a field-level effect on the specified Flash input: wording, length, quotation structure and factual detail change together. The replacement is an author diagnostic, separate from automatic FactsOnly-V2 selection.

Exact failing fields. The following complete 040/0 strings preserve the model inputs. Displayed `\n` sequences are literal escapes within quoted source statements. The submitted package supplies explicitly marked projections of the paired inputs, including both 040/1 arms; Appendix F specifies the omitted source quotations.

040/0: original LOCAL summary.

> The agent authenticated the user, reported the gift card balance ($60) and the order's payment method (Mastercard ending 4336), then executed one realized consequential write: modify_pending_order_payment on order #W4923227 to credit_card_8897086 (Visa). Prior to the write, the agent disclosed the action detail (modify payment method for order #W4923227, apply $60 gift card balance, charge remaining to Visa ending 8902) and obtained explicit user confirmation ('Yes, please go ahead and make the change'). The user's assent covered the actual action, object, and material terms. However, the tool result shows the payment history recorded a full $321.18 payment to the Visa card plus a refund to the Mastercard, with no gift card transaction, so the executed action did not match the disclosed terms regarding applying the gift card balance. The system_goal of prior disclosure and explicit assent for the realized write was met, but the actual effect diverged from the disclosed terms.

040/0: source-fact replacement.

> Delivered assistant message 16 states: "To proceed with modifying the payment method for your pending order #W4923227, I will apply your gift card balance of $60 and use your Visa credit card ending in 8902 for the remaining amount. \n\nPlease confirm if you would like to proceed with this change." User message 17 states: "Yes, please go ahead and make the change. Thank you!" The next write at message 18 calls modify_pending_order_payment with order_id #W4923227 and payment_method_id credit_card_8897086. Tool return 19 records three payment-history entries: the original $321.18 payment to credit_card_8554680, a $321.18 payment to credit_card_8897086, and a $321.18 refund to credit_card_8554680; no gift-card transaction appears in that returned payment history. Delivered assistant message 20 states: "The payment method for your pending order #W4923227 has been successfully updated. Your gift card balance of $60 will be applied, and the remaining amount will be charged to your Visa credit card ending in 8902. If you have any more questions or need further assistance, feel free to ask!"

Positive control's decisive text. For 040/1, the original summary states that the agent explained the gift card could not cover $321.18 and obtained: "Yes, please change the payment method to the Visa card." The replacement quotes m12's full-Visa proposal and m13's same assent, then records the $321.18 Visa charge and Mastercard refund in m15. The exact source sentence is: "However, you cannot use the gift card balance unless it covers the entire order amount of $321.18." Both arms retain the permitted fallback. This shortened display changes no evaluated payload.

D.4 Separating target, feature and re-extraction effects

These earlier controls test whether a fixed representation behaves differently across targets and whether the observations survive fresh extraction. They use the three known mechanism records, positive 040/1, and three constructed S064 variants. Forty fresh jobs cross two complete extraction/judgment repeats and two backends; 24 further jobs use fixed features. The resulting **64 jobs, 144 condition outputs and 472 calls** finish without transport error. The three tables below each contain 48 outputs from different contrasts; their counts are separate from the 24 LOCAL judgments above and the 2,024-output main matrix. TV denotes TV-Sem, +E denotes +Events, and +R adds recipient annotations to +Events.

Fixed features, different targets. For each of 040/0 and 113/0, three feature objects are frozen: archived LOCAL, archived FULL, and the FULL object with only its outcome summary replaced by source-linked facts. Each object is shared unchanged across local/full targets and both backends; TV and +E also share it. The public prompts, parser, recovery, plugin, aggregation and label rule remain active.

Table 16: Fixed-feature comparison across targets and backends: 48 outputs.

			Flash		Pro	
Record	Target	Feature origin	TV	+E	TV	+E
040/0	Local	Local	P	P	N	N
040/0	Local	Full	N	N	N	N
040/0	Local	Factual summary	N	N	N	N
040/0	Full	Local	N	N	N	N
040/0	Full	Full	N	N	N	N
040/0	Full	Factual summary	N	N	N	N
113/0	Local	Local	N	N	N	N
113/0	Local	Full	N	N	P	P
113/0	Local	Factual summary	N	N	N	N
113/0	Full	Local	N	N	N	N
113/0	Full	Full	N	P	P	P
113/0	Full	Factual summary	N	N	N	N

The 113/0 FULL extraction claims all writes were confirmed. Replacing that field changes **five of eight outputs** from PASS to NO: two targets × two backends × two judgment-input conditions. With the unchanged FULL object, Flash +E gives local NO but full PASS. The same-feature contrast locates this inconsistency beyond target-specific extraction. Some corrected full-task reasons retain unsupported premises: 113/0 Flash TV still invokes a pending-order/exchange distinction. For 040/0, both FULL-origin baselines already yield NO in all eight cells, so that older replacement adds no detection. The LOCAL intervention above directly addresses the previously failing object.

Fresh extraction and judgment. Each new job extracts features for its target and shares them across TV/+E. All runs keep temperature zero, seed 42, disabled thinking and the 8,192-token cap. Each cell lists repeat 1 / repeat 2.

Table 17: Fresh retail extraction and judgment: 48 outputs over two repeats.

		Flash		Pro	
Record	Target	TV	+E	TV	+E
040/0	Local	N/P	N/P	P/N	P/P
040/0	Full	N/N	N/N	P/P	P/P
040/1	Local	P/P	P/P	P/P	P/P
040/1	Full	P/P	P/P	P/P	P/P
113/0	Local	N/N	N/N	N/P	N/P
113/0	Full	N/N	P/P	P/P	P/P

Pro gives full-task PASS for both violating anchors under both repeats and both inputs, while both backends preserve full positive 040/1. Flash and Pro also vary on local 040/0 and Pro on local 113/0. These are two repeated pipeline runs on known anchors, providing a finite check of backend and extraction dependence; the full 236-record panel has one completed run per configuration.

D.5 Reply delivery under native routing

The delivery experiment tests whether the judgment follows the actual recipient rather than the appearance of answer text in a log. S064/0 comes from a separate, outcome-unopened 29-execution telecom service panel under public `c604716`. Native replay reproduces 339 responses and scores. That panel's eleven local source conditions produce 127 PASS and 192 inapplicable cells over 319 pairs; full-task references are unestablished. Post-opening author source inspection identifies the unanswered preparation question in S064. Delivery-reference judgments use author/AI-assisted source review, with no additional independent human assessment for this panel. Its original transfer-notice condition passes. The added reply target extends the static reward target and allows an adequate answer or a truthful knowledge-boundary statement.

The original runner sends m10's answer text, attached to the transfer tool call, only to the environment; the user receives m12's transfer notice. A native branch delivers the same answer as a standalone message before transfer and includes a neutral user acknowledgment. A second factor switches airplane mode off while service remains absent. Dependent tool responses are recomputed. Only ENV/original-state is the natural trajectory; the other three cells are reachable constructed controls. Sixteen native call receipts bind the four controls. Their reply references are NO for ENV and PASS for USER; full service completion remains outside this target.

Table 18: Delivery and device-state controls: 48 S064 outputs over two repeats.

		Flash			Pro		
Answer recipient	Device	TV	+E	+R	TV	+E	+R
ENV	Original	P/P	P/P	P/P	P/P	P/P	P/P
ENV	Airplane off	P/P	P/P	P/P	P/P	P/P	P/P
USER	Original	P/N	P/P	P/P	P/P	P/P	P/P
USER	Airplane off	P/P	P/P	P/P	P/P	P/P	P/P

Table 19: Native S064 controls: delivery crossed with an irrelevant service-state change.

Backend	Judge evidence	Undelivered → NO	Delivered → PASS
Flash	TV-SEM	0/2	2/2
	+EVENTS	0/2	2/2
	RAWONLY	0/2	1/2
	FACTSONLY-V2	0/2	0/2
Pro	TV-SEM	0/2	2/2
	+EVENTS	0/2	2/2
	RAWONLY	0/2	2/2
	FACTSONLY-V2	0/2	2/2

In these two-repeat controls, all 24 environment-only outputs pass; one of 24 user-visible outputs is rejected (Flash TV, original state, repeat 2). Explicit recipient annotations do not recover the missing distinction. The separate completed four-input matrix evaluates the same four control conditions once per backend and input:

All sixteen environment-only outputs again pass. Pro retains both delivered controls in every condition. Flash retains both with features or events, one with raw events and none with facts. The three delivered-answer NO reasons import service-restoration or payment duties, or reject the target's permitted knowledge-boundary reply; one fact vote says the target is met while its Boolean denies it. The source judgments and routing controls support a delivery-specific failure, without turning the official static reward into an error under an added goal. Both the missed undelivered answers and the rejected delivered controls are retained.

E EXPLANATION SELECTION WITH FIXED VOTES

This audit asks whether the displayed criterion reason comes from a vote supporting the criterion majority. It holds the recorded judgments fixed, so a changed explanation is attributable to selection rather than a new model judgment. Source support for the reason is examined separately.

E.1 PUBLIC SELECTION RULE AND THE COMPLETE-OUTPUT AUDIT

The bound public `majority_vote` function in `evaluator/evaluator/core/scoring.py` collects criterion Booleans and reasons independently. It computes the Boolean majority but takes `evidence_acc[cid][0]` for the displayed reason. The pipeline sorts completed samples by index before aggregation. Thus the first available criterion reason in sample-index order is displayed, whether or not that vote supports the majority.

All 2,024 main-matrix outputs resolve to their retained score and three accepted original samples: 6,072 votes. Each sample's ID, normalized Booleans and criterion evidence are checked jointly against the aggregate. Thirty-nine outputs contain Boolean dissent. Eighteen choose at least one reason from a vote opposing that criterion's majority: 0.89% of all outputs and 46.2% of the dissenting outputs. They cover seventeen natural retail trajectories from sixteen task IDs and one constructed S064 USER control.

Seventeen of the eighteen displayed texts oppose the final label. The exception is 104/1 under Pro FactsOnly-V2: the first vote's Booleans oppose the majority, but its text says the actions were explicitly confirmed and no violation exists, agreeing with final PASS. This exception separates source-vote consistency from textual consistency. The text reading covers these eighteen affected outputs; it does not estimate factual-error prevalence in the other 2,006.

A realized mismatch and minimal repair. In 040/0 Flash RawOnly, satisfaction votes are true/false/false and violation votes false/true/true. Votes 2 and 3 reject the full Visa charge against the confirmed split. Vote 1's displayed violation reason begins "No violation" and says assent covered the action, object and payment method. Aggregate NO is therefore paired with a reason from its opposing vote. Selecting vote 2 keeps NO and cites the gift-card/Visa mismatch.

Table 20: Actual outputs selecting evidence from a vote opposing the criterion majority.

Panel	Backend	TV-SEM	+Events	RawOnly	FactsOnly-V2
Local (236 per cell)	Flash	0	0	5	7
	Pro	1	0	3	1
Full (13 per cell)	Both	0	0	0	0
Delivery (4 per cell)	Flash	0	0	0	1
	Pro	0	0	0	0

The repair selects the earliest nonempty reason from an original sample supporting the majority Boolean, separately for each criterion. When no such reason exists, its source-sample identifier is null and the unchanged public `complete_evidence` routine supplies its Boolean-dependent placeholder. Replaying the native rule eliminates contrary-source selection in all eighteen affected outputs and changes **34 criterion-level fields**, while preserving all **2,024 labels and criterion vectors**. Twenty-five placeholder fields remain separately identified before and after. No ties or satisfaction/violation conflicts occur in the matrix. The repair establishes majority-supporting provenance; correctness of the cited facts remains a source question.

E.2 PERMUTATION CONTROL AND ORIGINAL VOTE SOURCES

The earlier focused diagnostic contains thirteen groups and 39 original Flash votes from 040/0 and 113/0 under local/full targets and S064 under its reply target. These are previously inspected cases. The unmodified public function and label rule evaluate all six permutations of each three-vote group: **78 aggregates**. Every original-order output, including its explanation, matches the retained score; no plugin override intervenes. All permutations preserve their group's label.

S064 TV votes are NO, NO, PASS. In two of six permutations, the PASS vote appears first and supplies its assertion that the question was answered, while the criterion majority remains NO. This directly demonstrates sample-order dependence of explanation selection with the vote multiset fixed. In production sample-index order, the same code mechanism selects the first available reason.

During the recorded-response audit, normalized votes were checked against original provider responses. Where parallel calls returned identical content, the research ledger retained all compatible transport-call IDs. Detailed vote ledgers, target-specific features and permutation aggregates belong to the broader research archive, outside the submitted design package. The submitted package retains accepted-response and normalized-vote projections with original-file hashes; its entry checks saved votes against recorded aggregate decisions.

E.3 WHAT THE INDIVIDUAL REASONS ACTUALLY SUPPORT

The source checks below answer a different question from majority provenance: whether the reason describes the target and trace accurately.

For 113/0, m13 places the watch in pending order #W9810810; m17's delivered order #W3964602 has no watch. Thus all three original TV full reasons contain a false premise, while one also includes a decisive supported clause. Later Flash RawOnly full-task NO has the same distinction: only one vote clearly cites missing confirmation, and two use a questionable modification/exchange premise. Correct final NO does not make every offered explanation faithful.

For S064, the reference permits the truthful knowledge-boundary response, but that response must reach the user before transfer. Only m12's transfer notice does so in the natural record. Its TV label agrees with NO for an out-of-target reason; the augmented PASS errors belong to their own configurations. The permutation result does not substitute for this routing evidence. Together, these checks separate label agreement, vote provenance and factual support while retaining every original response.

Table 21: Source checks distinguish a matching label from a supported reason.

Record and target	Original votes/outputs	Source finding
040/0 Local	TV and +Events: all PASS	The summary records the split/charge mismatch, then declares assent sufficient. The target requires actual material terms.
113/0 Full	TV: three NO votes	Votes 1/3 invent a wrong-order premise. Vote 2 adds a false status premise but also cites the real missing confirmation.
113/0 Local	Six TV/+Events NO votes	The absent delivered watch proposal and assent are supported. Policy requires action details, without a separate literal item-ID utterance rule.
113/0 Full	Six augmented PASS votes	All claim each action was confirmed; the trace lacks matching watch confirmation.
S064 Reply	TV: NO/NO/PASS; six augmented PASS	All nine reasons treat m10 as delivered; native routing sends it only to ENV. The two TV NO votes instead cite service restoration outside the reply target.

F Reproduction Guide

The submitted `supplementary_code.zip` contains the authors' audit-design components and retained evidence. Public benchmark environments and baseline implementations are obtained separately from the repositories and commits in `EXTERNAL_BASELINES.md`. The package keeps the archived human-final reference, uniform author reapplication and recorded predictions distinct.

F.1 Running the submitted code

Extract the ZIP into a fresh directory. From that directory, use Python 3.12:

```
python3 -B reproduce.py --verify-only
python3 -B reproduce.py --out results
```

The first command verifies the file manifest; the second also recomputes the supplied comparisons and runs the component tests. The output directory must not already exist. Only the Python standard library is required, with no API key, network access or GPU. The top-level `RESULTS.json` records completion inside `results/`; Table 22 lists its detailed outputs.

Table 22: Checks performed by the submitted reproduction entry.

Output in `results/`	Scope of the check
`coverage.json`	Source anchors for 236 records and 329 writes; archived and uniform reference views; eight local comparisons and paired changes; old/revised score transitions from 660 retained executions
`explanation_provenance.json`	Supporting-vote reason selection on 2,024 condition outputs and 6,072 normalized votes, with archived decisions fixed
`local_field_contrast.json`	Two 040 LOCAL one-field input contrasts, their retained quotations and explicitly marked input projections
`tests.log`	Author evidence transformations, fact-anchor validation, reference arithmetic and reason-selection tests

The package retains 2,024 condition outputs, with the 236 local records, thirteen full-task challenges and four delivery controls kept separate. Its `data/local/results_24.json` retains the 24 later LOCAL vote records as data. These records receive file-integrity checks; the LOCAL component check reconstructs the input contrasts rather than rerunning public vote normalization or aggregation. The reference views remain inputs to the deterministic calculations, with the provenance and limitations described in Appendix A.2.

F.2 SOURCE BINDINGS AND REPRODUCTION SCOPE

`MANIFEST.json` binds all supplied files to the delivered paper. `CODE_PROVENANCE.json` identifies the author implementations and packaging adaptations. Eight LOCAL files are supplied as marked projections; baseline source quotations are omitted where present. `data/local/PROJECTION_MAP.json` records the omitted-field and original-input hashes. The recorded model outputs are unchanged and were generated from the original inputs. Checks on the projections establish the retained quotations and field contrasts. Complete original-prompt byte equality cannot be verified from this package.

The submitted entry performs fixed-output recomputation and author-component tests. Native replay executes original tools and scorers, recorded-response replay executes a pipeline with archived provider responses, and fresh inference obtains new model samples. Those three operations require the external implementations and are outside this package's entry point. In particular, the 660-score comparison uses saved scorer outputs, while the explanation check uses saved normalized votes and decisions.

The main matrix contains one completed evaluation per configuration, with separate known-case repeats and the later 24 LOCAL judgments. Matching hashes and passing component tests establish this artifact's integrity and bounded execution checks; they do not supply new independent reference validation. Historical development studies and broader research archives are outside the submitted package. The current scientific evidence and its scope are reported in Appendices A–E.