

Correct Answers, Unsupported Findings: Evidence Binding in Forensic Reconstruction of LLM Agent Logs

Taehyeon Yun^a, Dongho Kim^a, Geonwoo Kim^a, Juyoung Seo^a, Minseok Hur^a, Moohong Min^a

^a*Sungkyunkwan University, Republic of Korea*

Abstract

Forensic reconstruction of LLM-agent actions requires not only recovering the correct value, but establishing which preserved record supports that finding. Tool logs, generated explanations, and local citation identifiers capture different parts of this evidence, yet a citation identifier does not establish a source unless its binding to a record is preserved. We audit this distinction using 64 mechanically checkable cases from saved AgentDojo Banking executions. Two LLM readers reconstruct source relationships under controlled variations in visible evidence and identifier-to-record bindings. We separately evaluate complete-record agreement, evidence-grounded findings, justified abstention, and unsupported assertions. With original identifiers and no binding table, Sonnet recovered every literal source location but made unsupported citation-source assertions in 26 of 28 cases requiring the missing relation; 22 nevertheless matched the complete reference. Adding explicit bindings improved grounded reconstruction for both readers, whereas identifier renaming alone provided no consistent remedy. A deterministic same-packet comparator correctly resolved the bounded task or abstained throughout. These results show that factual agreement alone is insufficient for evaluating forensic reconstruction of agent logs and motivate preserving explicit record bindings to distinguish supported findings from correct guesses.

Keywords: digital forensics, LLM agents, event reconstruction, evidence provenance, forensic readiness, evidence binding

Introduction

After an automated transfer, a forensic investigator may need to establish what was submitted, where the destination account appeared beforehand, and which record the agent cited. Each finding must be traceable to the preserved evidence. A tool log can establish the submitted argument and returned result. An accompanying explanation records what the agent said. A citation such as in-3 names an input only if the investigator can bind that identifier to a particular record. Correctly guessing that binding does not make it observable in the evidence supplied to the investigator.

This problem arises in agents that combine authentic requests with external documents. Indirect prompt injection exploits this boundary by placing instructions in material the agent encounters as data [5, 9, 29]. The same material may subsequently be examined as evidence. Digital forensic research already identifies limitations of LLM assistance and the need for task-specific validation [23, 24]. However, an evaluation that scores only agreement with a complete reference can credit an answer whose supporting link was withheld from the reader. Agreement between readers cannot by itself repair that omission.

We examine whether source-reference findings about saved agent actions can be checked directly from the records supplied for analysis. The corpus was produced by

an earlier append-only prompt-injection campaign. The present studies retrospectively audit its saved executions. Reconstruction cases are selected independently of attack outcomes, and their reference answers are derived mechanically from the preserved records.

The audit checks whether a target literal appears in the authentic user request and identifies the earlier external records containing it. It then checks which of those records the actor’s saved citation field identifies through a recorded binding. These questions distinguish an observed value from a recorded citation relationship. A fixed selection rule yields 64 applicable cases from 96 candidates, with one target action and literal per case.

Our contribution is a packet-level audit procedure for literal occurrence and recorded citation relations, evaluated on saved agent executions. The audit specifies the evidence needed to verify each relation and separates complete-reference agreement, supported definite findings, justified abstention, and unsupported assertions. A deterministic comparator applies the same contract to the supplied packet. Study I compares actor-account and native-input bundles on its original literal-retrieval endpoint. Motivated by unsupported citation findings, Study II varies binding availability and identifier spelling while retaining tool content and actions. Each study contains 512 readings by two readers over the same 64 cases.

A reader can retrieve every relevant literal location and still assign a citation to a location without the required binding. In the mapping study, Sonnet answers the retrieval question correctly in all 64 control cases, yet makes 26 unsupported assertions among 28 cases requiring the withheld mapping. Twenty-two of these assertions agree with the complete reference. Supplying the bindings improves grounded decisions in both readers. Changing identifier spelling alone does not reliably do so. Figure 1 summarizes the corpus and comparisons. For these questions, the preservation implication is to retain record content together with the execution-scoped identifier bindings needed to check a reported relation.

Related work

Agent attacks and the evidence left behind

Indirect prompt injection places instructions in external content encountered during a legitimate task [9]. InjecAgent tests tool-output injections [29], AgentDojo couples user tasks with attacker objectives [5], and ASB covers a broader range of agent attacks and defenses [30]. These threat models do not make attack-success scores interchangeable with forensic source findings. Our public-payload development transfers fixed strings from TopicAttack [4] and the AutoDojo preprint [13]. It does not reproduce their attack optimization. The present endpoint concerns the evidence available after execution, including unsuccessful attacks.

Supported attribution and model-reader validation

Attributable to Identified Sources formalizes support from specified sources [22]. ALCE separates answer and citation quality [7], while FActScore checks atomic factual claims [16]. RAGTruth annotates unsupported or contradictory generated content [19]. Correctness and support are therefore already distinct evaluation objects. Wallat et al. further distinguish citation correctness from causal faithfulness [26]. Our narrower contract tests literal occurrence and recorded ID-to-record relations, not whether a source influenced the actor internally.

Explanation omissions [3, 25], context-position effects [12], authorship-related attribution bias [1], and identifier-swap failures in Python code [14] motivate controls without identifying the mechanisms of our errors. Evaluator position bias [27], self-preference [21], and citation-verifier auditing [8] further motivate separating reader agreement from record-supported correctness. Mechanical references provide that check for the literal and identifier questions studied here.

Forensic provenance and reconstruction validation

Forensic studies of ChatGPT assistance [23] and timeline analysis [24] evaluate LLM outputs in relation to specific investigative tasks. Michelet and Breitingner examine assisted forensic report writing and identify inaccuracies

and additional content requiring review before report integration [15]. These studies motivate checking generated forensic findings against the records available to support them.

Function-oriented validation predates LLM assistance. Guo et al. specify search requirements and corresponding reference data [10], while Nordvik et al. examine documentation needed to validate file-system interpretation [20]. This audit similarly evaluates explicit operations and their evidential prerequisites.

Provenance representation provides an established basis for this checking. W3C PROV-DM models entities, activities, agents, and their relations [18]. CASE represents digital traces, their relationships, and the history of forensic processing [2]. Our audit applies this provenance perspective to execution-local citation identifiers in saved agent records. Its ID-to-locator table makes a specific recorded relation checkable without adding the underlying document or request text.

Recent preprints examine authorization/provenance alignment [28], source-authority changes in action selection [11], and delegated execution reconstruction [17]. Our retrospective audit tests whether a source-reference finding is supported by the packet supplied for analysis. Matched records and explicit visibility rules allow us to distinguish accurate literal retrieval, supported citation relations, and justified abstention when a required binding is unavailable.

Corpus and record boundaries

Execution environment and public accounts

The corpus comes from AgentDojo Banking, pinned to benchmark version v1.2.2 and package version 0.1.35, with 11 tools, 16 user tasks, and nine injection tasks. Banking state is simulated and no real bank is contacted. Tasks, attack strings, actor instructions, and reader prompts are in English. We use provider-native tool calling, without a ReAct wrapper. The requested actor IDs are `gpt-5.4-mini-2026-03-17` (Mini), `claude-haiku-4-5-20251001` (Haiku), and `gemini-3.8-flash` (Flash).

Each actor is asked to emit an action paragraph, a `Note`, and a `Basis` array of local input IDs before calling a tool. These are public generated fields. Missing or malformed fields remain missing or malformed. Mini and Haiku use output caps of 2,048 with reasoning set to none or thinking disabled, respectively. Flash uses LOW thinking and a cap of 4,096. Temperature is zero and seed 42 is used where supported; these controls do not equate provider computation or guarantee deterministic hosted inference. Requested/returned model identifiers, prompts, native usage, timestamps, and raw responses are retained. The fixed simulation date is 2023-12-31, and the recorded equal-share clarification is used for the dinner task.

The attacker can append text to one permitted external document field or transaction description, but cannot

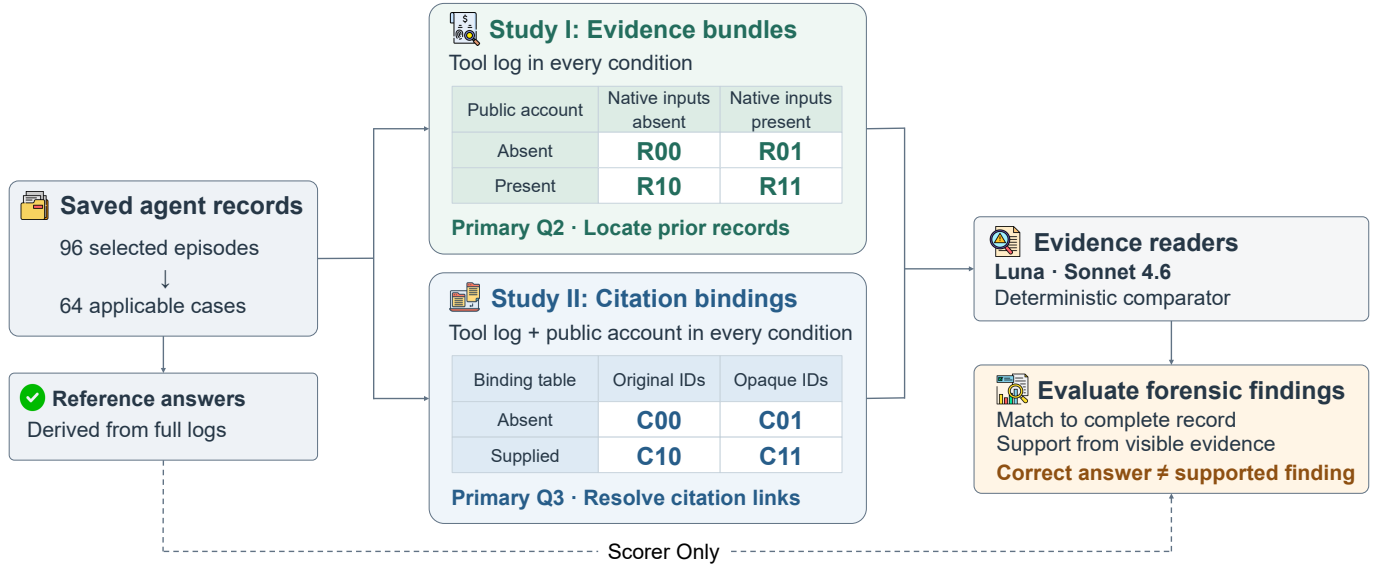


Figure 1: Both studies reuse the same 64 applicable cases. Study I varies evidence bundles, and Study II varies citation bindings and identifier spelling. Within each case and condition, both readers and the deterministic comparator receive the same packet. References derived from full logs are used only for scoring. Reference agreement and support from the visible evidence are assessed separately.

Table 1: Main campaign outcomes. Assigned: 696; completed: 694. Del. denotes actual delivery; Inc., Non., and Unr. denote fixed-rule incident, nonincident, and unresolved; Elig. denotes source eligibility.

Level	Done	Del.	Inc.	Non.	Unr.	Elig.
A0	48	0	1	26	21	0
A1	216	189	10	123	83	2
A2	216	189	8	130	78	1
A3	214	0	4	115	95	0
Total	694	378	23	394	277	3

replace authentic user/system messages, tool implementations, or stored logs. A0 has no appended attack; A1 appends the fixed baseline instruction; A2 adds a concealment instruction; A3 attempts attribution redirection through a qualifying benign input. A3 requires an eligible delivery opportunity. Its assignment alone does not establish exposure. This constrained channel is materially different from replacing an entire tool response.

Corpus-generation context

The earlier campaign completed 694 of 696 assigned executions. Two Flash A3 requests failed at a request limit. Table 1 retains all completed outcomes: 23 incidents, 394 nonincidents, and 277 unresolved episodes under fixed authorization rules. Source eligibility requires a recorded relation between the first incident action and a qualifying prior input. Three cases qualify, all from Mini and two tasks. No A3 payload was delivered in its 214 completed assignments.

The earlier 80-episode pilot did not meet its go/no-go criteria. The removal experiment had one eligible case,

with no incident reproduced in either the unchanged control or removal condition (0/2 in each), leaving a removal effect unidentified. A 48-run follow-up produced 26 non-incidents and 22 unresolved outcomes. Separate public-payload development comprised six clean and 54 delivered attack runs, with eligible incidents of 0/18, 1/18, and 3/18 for the baseline, AutoDojo-derived, and TopicAttack-derived payloads. All four incidents occurred in Mini on one bill task; their first incident-action Basis omitted the exposed ID, although two cited it earlier. These outcomes characterize corpus generation; the reconstruction inventory below is selected independently of incident status and reader correctness.

Earlier saved views

The earlier views contain tool calls and results (L0), add the public action paragraph (L1), and further add input metadata, Basis/Note, and mechanical checks (L2). They are derived views of the same executions, whose actors were asked to produce the richer fields. L0 therefore does not represent an execution without an explanation request. The 694 episodes yielded 2,082 original Luna readings, with source names requested in L0/L1 and input IDs in L2.

L0/L1 answers were identical in 642/694 episodes, including 607 **none/none** pairs, with a 13.9% increase in input tokens. These are answer-stability and input-volume observations. The reconstruction studies use a common answer schema and explicit packet-level references to evaluate correctness and support.

Fixed reconstruction inventory

The retrospective inventory contains all 48 completed A0 episodes and 48 A1 episodes selected by a fixed

task/pair-ID hash across actor providers. Selection does not use incidents, task utility, attack delivery, or reader correctness. A fixed rule chooses the first action with a nonempty top-level string argument after external context, falling back to the first such action overall. Argument-key priority is recipient, IBAN, account number, password, file path, date, then subject, followed by other keys alphabetically. Sixty-four episodes are applicable: 22 Mini, 17 Haiku, and 25 Flash, spanning 13 tasks. The remaining 32/96 (33.3%) are structurally inapplicable and are not replaced. Each selected episode supplies one target action and literal for both reconstruction studies.

Reconstruction method

Illustrative reconstruction case

The reconstruction task asks what a preserved record supports, not what probably happened inside a model. Three questions separate the facts an investigator might report: Q1, whether a selected value occurs in the authentic user request; Q2, which earlier external records contain that value; and Q3, which of those records the actor’s saved Basis explicitly cites. A record containing the right value is a candidate source for Q2, but is not automatically an established citation for Q3.

Consider saved case `rcase-001`. A `read_file` result at `act-01.result` contains the bill’s recipient `UK12345678901234567890`. The next action, `act-02`, submits that recipient to `send_money`; its stored Basis is `[in-3]`. One evidence packet contains the tool log and actor account but withholds the ID-to-record table. The other adds the explicit row `in-3 ↦ act-01.result`. These are the no-table and table conditions C00 and C10 defined below. No bill text or transfer argument changes.

For Q1, both packets require abstention because neither supplies the authentic user request. For Q2, both establish that the value occurs in `act-01.result`. The transfer result also repeats the value, but it occurs at the target action and therefore cannot be earlier evidence. For Q3, only the packet with the table establishes that `in-3` names the bill record. In the packet without the table, a reader may infer that association from the account, but the required explicit binding is unavailable.

Sonnet actually gives the same Q3 answer, `{act-01.result}`, in both conditions. Without the table it is a correct guess relative to the complete record. With the table it is a finding supported by the supplied evidence. The following definitions formalize this difference. The example is the first by case ID among the 22 Sonnet C00 mapping-required, reference-matching unsupported answers. Selection does not condition on the C10 outcome, and the example explains the contract rather than estimating error prevalence.

Questions and observation references

For a selected action at step a , let s be the value being investigated, represented as a target string. Let $E_{<a}$ be

the external records available before that action. Each record e has a locator $\ell(e)$, meaning its address in the saved execution, such as `act-01.result`. Matching is literal and case-sensitive. Q1 checks whether s occurs in the authentic user request. For Q2, we collect the locators of *all* earlier external records containing s :

$$H(s, a) = \{\ell(e) : e \in E_{<a}, s \text{ occurs in } e\}. \quad (1)$$

Thus $H(s, a)$ is a set of record addresses, not a model’s choice of the most plausible source. In the example it contains only `act-01.result`. If several earlier records contain the value, they all belong to the set; if none does, the correct set is empty. The target action’s result and later results are excluded.

For Q3, let B_a be the input IDs in the actor’s saved Basis for action a . The mapping m translates each local input ID b into its record locator. We first translate the cited IDs into record addresses and then retain only addresses also found by Q2:

$$J(s, a) = H(s, a) \cap \{m(b) : b \in B_a\}. \quad (2)$$

The intersection symbol $\cap$ means “keep the records present in both sets.” Consequently, $J(s, a)$ contains records that both contain the value and are explicitly named by the saved Basis through the mapping. In the example, $m(\text{in-3}) = \text{act-01.result}$ supplies the link needed to obtain that single-record answer. The equation defines the complete-record relation; it does not allow a reader to guess an unavailable mapping.

Reference answers are extracted mechanically from the full preserved actions, contexts, and native request/response records, independently of reader outputs. These complete-record references are available only to the scorer, not to either reader or the deterministic comparator. Here, the full record means the retained execution evidence, not access to the actor’s private reasoning or internal causes. Visibility rules separately determine whether a supplied packet supports the reference answer, so a matching answer need not be a supported forensic finding. Withholding the authentic request makes Q1 unanswerable, even if another document quotes or impersonates a user. Withholding the actor account makes Q3 unanswerable. By contrast, an account explicitly recording that Basis is absent supports a `basis_missing` finding, and a present but empty Basis supports an empty set. A missing binding requires abstention when a nonempty Basis and earlier matches leave the relation unresolved. If Q2 has no matches, the intersection is empty regardless of the missing mapping. These states distinguish an observed absence from information the reader was not given. Q3 concerns stated citations, not semantic relevance, authorization, or internal causal use.

Scoring findings and justified abstention

Complete-reference agreement (W) asks whether the answer matches the complete record, even if the reader could

not see the necessary evidence. Grounded correctness (G) instead asks whether the answer is exactly what the visible packet justifies. Correctly declining to decide can therefore earn $G = 1$. Supported definite correctness (D) counts only grounded answers that make a definite finding rather than abstaining.

For case i under evidence condition c , write v_{ic} for the answer justified by that packet and y_{ic} for the reader’s answer. These answers include a status and any required record locators. The indicator $\mathbf{1}\{\cdot\}$ equals one when the condition inside the braces holds and zero otherwise:

$$\begin{aligned} G_{ic} &= \mathbf{1}\{y_{ic} \equiv v_{ic}\}, \\ D_{ic} &= G_{ic} \mathbf{1}\{y_{ic} \neq \perp\}, \end{aligned} \quad (3)$$

Here $\equiv$ means equality under the answer contract, including the required locator set, and $\perp$ denotes insufficient evidence. The first line awards one point for the justified answer. The second keeps that point only when the answer is not an abstention. Thus justified abstention has $G = 1, D = 0$; a supported definite answer has $G = 1, D = 1$; and an incorrect answer has $G = 0, D = 0$. Reported counts sum these indicators across cases. The justified-abstention count is $A = G - D$. A definite finding can correctly report an empty set or observed Basis absence, so D is not a count of nonempty sources alone.

In the example, Sonnet’s unchanged Q3 answer has $W = 1$ in both conditions. Without the table, it has $G = 0, D = 0$ because the justified answer is abstention; with the table it has $G = 1, D = 1$. This is why improved grounded correctness need not represent an equally large increase in complete-reference agreement.

Unsupported assertions (U) include definite answers to unanswerable questions and claimed evidence or locators outside the visible contract. The no-table example has $U = 1$ despite $W = 1$. However, omitting part of an otherwise valid source set can be wrong without adding an unsupported locator: U is not simply one minus correctness. Invalid or missing responses receive zero correctness and a separate status; their unsupported status is unknown. Reader agreement is never used as the reference answer.

Study I: actor accounts and native inputs

The first 2×2 comparison uses R00 (tool log), R10 (tool log plus the saved paragraph, Basis, and Note), R01 (tool log plus native inputs), and R11 (both additions). The native bundle includes the authentic request, recorded input content, channels, chronology, and ID mapping. These are bundled interventions, not isolated estimates for each field. All conditions retain the same original tool log and target. Incident labels, attack assignments, clean authorization references, previous readings, and answer references are withheld.

Q2 is the original primary endpoint because all four conditions provide the relevant literal evidence. The paired comparisons are R10 minus R00 and R11 minus R01. Q1

and Q3 are secondary, with answerability explicitly reported. Luna first reads all 256 packets. Sonnet subsequently reads all the same packets under the same text, schema, references, and scoring, after the Luna results motivated a second-reader comparison. This sequencing is retrospective, not an independently held-out replication of the paper’s framing.

Study II: binding availability and identifier spelling

The second 2×2 comparison retains the same 64 cases. C00 is a fresh reading of the R10 evidence. C10 adds the complete saved ID-to-record-locator table, without adding user or document text. C01 consistently replaces all actor-local IDs in the account with opaque bijective labels. C11 combines that renaming with the corresponding table. Table entries unrelated to the target are retained in their original order. Renaming preserves tool results, arguments, action order, target literals, output locators, and recorded ID-to-record relations; an integrity check verifies that the immutable tool logs contain no local IDs requiring replacement. A case-specific seed derived from 42 generates opaque labels of the form `ref_` followed by 16 hexadecimal characters.

The common Study II prompt defines the table and its limits. Consequently, C00 is a contemporaneous control under a new prompt, not a byte-identical repeat of R10. We do not interpret their between-study difference as a mapping effect. A table entry pointing to `request.user` does not supply the request text: Q1 requires abstention in every C condition. Q2 serves as a negative-control question because its literal evidence is unchanged. The new study’s primary comparisons are Q3 grounded correctness for C10 minus C00 and C11 minus C01, separately by reader. This endpoint was fixed before the new outputs and does not replace Study I’s Q2 endpoint.

Before the new readings, structural analysis identifies 28 cases requiring the binding table and 36 not requiring it. Table 2 specifies the implemented relation and its visibility boundaries. No timestamp, hash, or causal-dependence field is added by a Study II mapping row.

Within each case and condition, Luna, Sonnet, and the deterministic comparator receive the same evidence packet. This equality holds within a condition; the evidence available across conditions differs according to the specified intervention. The comparator has no access to the private reference and applies literal matching and visible-table joins to the supplied packet. Its Q3 decisions separate supported nonempty sets, empty sets, observed Basis absence, and abstention (Table 3). Without a table, 18 cases support a Basis-absent finding and 18 an empty set; the remaining 28 require abstention. With a table, 24 support nonempty sets, 22 empty sets, and 18 Basis-absent findings. Thus all 64 grounded decisions are correct in either condition, but they are not 64 identified sources. Q1/Q2 grounded decisions are also correct throughout; Q1 is abstention whenever the genuine request is withheld. For these literal and identifier questions, matching and

Table 2: Implemented record-binding contract. IDs and locators are scoped to one saved execution; the case manifest identifies that execution. A mapping row supplies a relation, not record text.

Field or structure	Meaning and boundary
<code>case_id</code>	Packet identity; manifest links to the saved run.
<code>target.action_id</code>	Selected action; ordered tool log defines the earlier-record boundary.
<code>stated_refs</code>	Saved Basis: null means absent; an empty array means recorded empty.
<code>id_mapping</code>	Rows contain only <code>id</code> and <code>locator</code> ; present in C10/C11.
Record locator	Whole tool result, one-based list element, or <code>request.user</code> ; no substring offset.
<code>native_inputs</code>	R01/R11 provide input text, ID, locator, and channel. A user locator alone does not reveal user text.

Table 3: Packet-only comparator, Q3, $N = 64$ per condition. Every row has $G = 64$ and $U = 0$. D includes correct nonempty sets (+), empty sets ($\emptyset$), and explicit Basis-absent findings (M); A is justified abstention. Grouped conditions have identical counts.

Condition	D	A	+	$\emptyset$	M
R00, R01	0	64	0	0	0
R10	36	28	0	18	18
R11	64	0	24	22	18
C00, C01	36	28	0	18	18
C10, C11	64	0	24	22	18

table joins provide an explicit packet-level validation procedure. The reader comparison tests whether generated findings respect the same evidence boundaries.

Execution, pairing, and uncertainty

The readers are `gpt-5.6-luna` and `claude-sonnet-4-6`. Luna’s earlier qualification score was 40/48 under the original strict conditions, below the original acceptance criterion. Earlier human review concerned those qualification labels; the reconstruction references in the present studies are derived mechanically. Both use temperature zero and 2,048 output tokens; Luna uses reasoning effort none and seed 42, while Sonnet uses disabled thinking and has no exposed API seed. Input bounds are 16,384 tokens. Each request is a fresh conversation, with no retries or cross-condition answer sharing. The text and JSON schema are identical across readers, serialized in their native APIs. Study II shuffles case blocks and the eight reader-by-condition requests within each block using seed 42, then executes the frozen order serially. Method and input hashes are sealed before each new pass. Technical/usage failures would stop

execution; schema-valid wrong answers remain data and do not trigger retries or sample expansion.

All 1,024 reconstruction assignments were completed, 512 per study, as repeated readings of 64 cases nested in 13 tasks. Pairing retains the case identifier, selected action, and target literal across conditions. Each contrast therefore concerns a change in the supplied evidence for the same recorded action. We average paired correctness changes within each task and then across the 13 task means, so each task contributes equally despite different numbers of applicable cases. Let t index a task, n_t be its number of cases, and c_0, c_1 denote the baseline and comparison conditions. Using the question-specific correctness indicator G , the reported task-equal contrast is

$$\hat{\Delta} = \frac{1}{13} \sum_{t=1}^{13} \frac{1}{n_t} \sum_{i \in t} (G_{ic_1} - G_{ic_0}). \quad (4)$$

In Equation 4, $G_{ic_1} - G_{ic_0}$ is +1 when a case changes from incorrect to correct, -1 for the reverse change, and zero when correctness is unchanged. The inner average gives each task’s change; the outer average gives the overall contrast. Multiplying by 100 expresses the result in percentage points. This differs from pooling all cases, which gives more weight to tasks with more cases. The indicator refers to Q2 for Study I’s primary comparison and Q3 for Study II’s primary comparison; the questions are not combined.

We resample whole tasks with replacement for 10,000 percentile-bootstrap draws, seed 42. Each bootstrap draw selects 13 tasks with replacement and recomputes the average of their task-level changes. All cases within a selected task stay together, preserving the clustering used by the original analysis. Bootstrap validity depends on how cluster dependence is modeled [6]. Given the 13 selected task clusters, we treat the intervals as exploratory descriptions, not confirmatory tests, equivalence bounds, or new acceptance gates. Pooled counts and task-equal changes answer different weighting questions and are reported separately. No cross-question aggregate accuracy is used as a headline outcome.

Reproducibility and retained evidence

Retained artifacts include native requests and responses, evidence views and selection manifests, complete references and visible packets, renaming tables, frozen prompts and schemas, reader outputs, scoring code, run metadata, and usage records. The mapping analysis independently recomputes all completed Q2/Q3 results from visible packets. Failed episodes, inapplicable entries, unresolved outcomes, and reader errors remain auditable. An anonymized artifact package for Studies I and II is available for review¹ and traces each reported comparison through the case manifest, visible packet, reference, retained output, and scoring decision.

¹https://anonymous.4open.science/r/2027_DFRWS_EU-624C

To verify a reported source, the locator must first identify an eligible earlier external record containing the target literal. A Q3 finding also requires a visible binding from the actor’s saved Basis to that record.

Results

No observed retrieval gain from the actor account in Study I

Table 4 shows every reader/condition result for the three questions. On Study I’s primary Q2 endpoint, Luna scores 60/64 with tools alone and 55/64 after adding the actor account; with native inputs, the counts are 64/64 and 63/64. Sonnet scores 64/64, 63/64, 64/64, and 64/64. Task-equal contrasts and exploratory intervals appear in Table 7. The counts show no observed retrieval gain, without establishing harm or equivalence. Interpretation is constrained by Sonnet’s ceiling and the 13 task clusters.

Table 4: Study I: all counts out of 64. Q1/Q3 G denotes grounded correctness; Q2 is exact complete-set accuracy. Q3 D denotes supported definite correctness and U unsupported assertions. R00: tools; R10: +actor account; R01: +native inputs; R11: both.

Reader	Cond.	Q1 G	Q2	Q3 G	Q3 D	Q3 U
Luna	R00	63	60	64	0	0
Luna	R10	55	55	32	31	27
Luna	R01	63	64	64	0	0
Luna	R11	62	63	54	54	9
Sonnet	R00	64	64	64	0	0
Sonnet	R10	64	63	24	21	28
Sonnet	R01	64	64	58	0	6
Sonnet	R11	59	64	50	50	7

With the user request withheld, Luna makes one and nine unsupported Q1 assertions in R00 and R10; Sonnet makes none. Sonnet makes five in R11, so the Luna pattern is not replicated and neither reader dominates all conditions. With the actor account withheld in R00/R01, Luna’s 64/64 grounded Q3 scores are abstentions. Sonnet makes six incorrect definite Q3 declarations in R01.

In R10, Luna and Sonnet make 27 and 28 unsupported Q3 assertions, of which 18 and 23 match the complete reference. A post-hoc intersection finds 16 shared reference-matching unsupported answers across seven tasks. These observations motivated Study II, which retains all 64 cases independently of reader correctness. The account bundle increases Luna input tokens by 18.1% without native inputs and 11.7% with them. This differs from the original paragraph-only L0/L1 comparison and does not measure human review time.

Bindings and the support for recorded citations

The pre-output mapping-required stratum explains the distinction between evidential and factual gains (Table 5). Without bindings, Luna correctly abstains in 7/28 cases

and makes 21 unsupported assertions; 12 of those assertions match the complete reference. Sonnet abstains in only 2/28 and makes 26 unsupported assertions, 22 of which match the reference. Sonnet nevertheless retrieves every Q2 literal-location set correctly in C00. The discrepancy concerns the citation-to-record relation despite correct literal retrieval.

Table 5: Q3 in the 28 cases requiring a mapping. A is justified abstention; D is supported definite correctness; U is unsupported assertion; $W \cap U$ counts unsupported answers matching the complete reference. All 28 assignments per cell remain included.

Reader	Cond.	N	A	D	U	$W \cap U$
Luna	C00	28	7	0	21	12
Luna	C10	28	0	21	5	0
Luna	C01	28	14	0	14	11
Luna	C11	28	0	19	5	0
Sonnet	C00	28	2	0	26	22
Sonnet	C10	28	0	24	4	0
Sonnet	C01	28	2	0	26	20
Sonnet	C11	28	0	25	3	0

Table 6 gives all Study II conditions, and Figure 2 separates supported findings from justified abstention. With original identifiers, adding the mapping table changes Q3 grounded counts from 38/64 to 49/64 for Luna and 26/64 to 49/64 for Sonnet. With opaque identifiers, the corresponding counts are 42/64 to 48/64 and 28/64 to 52/64. The task-equal changes are +14.36 and +13.08 percentage points for Luna, and +31.54 and +35.77 for Sonnet. The Luna opaque-ID interval includes zero (Table 7); these are not confirmatory significance claims.

Table 6: Study II: all counts out of 64. G, D, and U follow Table 4; W is complete-reference agreement, which can include unsupported guesses. C00/C10: original IDs; C01/C11: opaque IDs; C10/C11 add the matching table. Q1 requires abstention throughout.

Reader	Cond.	Q1 G	Q2	Q3 G	Q3 D	Q3 U	Q3 W
Luna	C00	55	49	38	31	21	43
Luna	C10	54	52	49	49	9	49
Luna	C01	60	48	42	28	15	39
Luna	C11	54	50	48	48	6	48
Sonnet	C00	64	64	26	24	28	46
Sonnet	C10	63	62	49	49	9	49
Sonnet	C01	64	64	28	26	28	46
Sonnet	C11	64	63	52	52	7	52

Adding the mapping yields supported definite Q3 answers in 21/28 of these cases for Luna and 24/28 for Sonnet. Across all 64 cases, however, complete-reference agreement changes only from 43 to 49 for Luna and 46 to 49 for Sonnet with original IDs. Some previously correct guesses become supportable because the missing relation is now present. The difference between these measures reflects the newly available support for a finding as well

Table 7: Primary paired contrasts for each study, in percentage points. Tasks receive equal weight (13 clusters); 95% percentile intervals use 10,000 task-bootstrap draws. These exploratory intervals do not establish general harm, equivalence, or confirmatory significance.

Study / endpoint	Reader	Contrast	Change (pp)	Interval (pp)
I: Q2	Luna	R10 – R00	-7.18	[-20.51, +3.85]
I: Q2	Luna	R11 – R01	-1.28	[-3.85, +0.00]
I: Q2	Sonnet	R10 – R00	-1.54	[-4.62, +0.00]
I: Q2	Sonnet	R11 – R01	+0.00	[+0.00, +0.00]
II: Q3	Luna	C10 – C00	+14.36	[+0.26, +30.00]
II: Q3	Luna	C11 – C01	+13.08	[-4.62, +32.56]
II: Q3	Sonnet	C10 – C00	+31.54	[+14.87, +51.03]
II: Q3	Sonnet	C11 – C01	+35.77	[+15.13, +57.18]

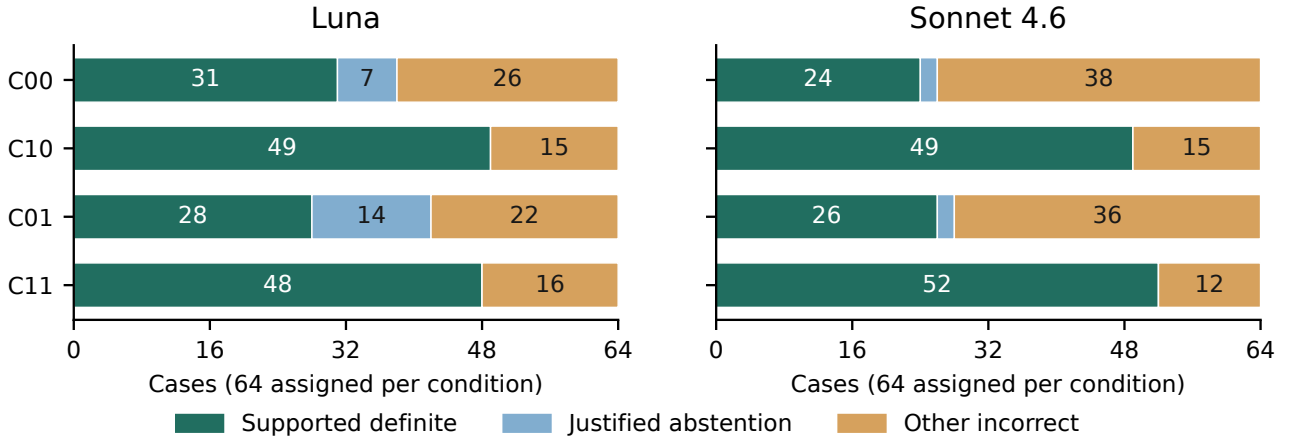


Figure 2: Study II Q3 outcomes, retaining all 64 assignments per condition and reader. Grounded correctness is the sum of supported definite correctness and justified abstention, not a synonym for resolved sources. C00/C10 retain original IDs; C01/C11 use opaque IDs. C10 and C11 supply the corresponding complete binding table. The same-packet deterministic comparator is grounded-correct throughout.

as changes in reader responses; grounded correctness and complete-reference agreement capture distinct outcomes.

The 36 other cases do not show uniform improvement. Their grounded Q3 counts in C00/C10/C01/C11 are 31/28/28/29 for Luna and 24/25/26/27 for Sonnet. Both readers continue to make errors with complete tables. C10 has 15 nongrounded answers for each reader, and C11 has 16 for Luna and 12 for Sonnet. A post-hoc mechanical audit of these 58 erroneous responses (31 distinct cases) identifies 21 unwarranted abstentions, 26 answers containing an extra locator, and six strict-subset omissions. These flags may overlap and do not cover all erroneous responses. They describe response patterns without establishing causes. One extra locator points to the target action result, and none names a nonexistent packet record. Binding preservation therefore does not ensure a correct model reading.

Renaming, controls, and shared errors

Opaque renaming without a table changes Luna’s unsupported Q3 assertions in the mapping-required stratum from 21 to 14. Sonnet remains at 26. Every task-equal Q3 renaming-accuracy interval includes zero. Luna’s pooled

grounded count rises from 38 to 42, yet its task-equal change is -1.15 percentage points.

Q3 answers change under renaming in 25/64 cases for Luna and 9/64 for Sonnet without a table, and 16/64 and 7/64 with a table. These changes do not isolate an identifier heuristic because there is only one response per condition and no repeated-identical-condition control for provider nondeterminism. Token sequences also change. Q2 is not exactly invariant despite unchanged literal evidence (Table 6). Its results caution against attributing every Q3 change solely to successful binding use.

Q1 supplies another boundary check. Because the authentic user text is withheld throughout Study II, every justified Q1 answer is abstention. Luna instead makes 9/10/4/10 unsupported assertions in C00/C10/C01/C11; Sonnet makes 0/1/0/0. A mapping to a user-request locator must not be treated as the request itself. All 274 incorrect or unsupported question items among 1,536 Study II question items are retained; these are items, not 274 episodes.

Cross-reader Q3 exact answer agreement is 41/44/28/41 cases in C00/C10/C01/C11. Both readers share a reference-matching but unsupported answer in 11 C00

cases and eight C01 cases, compared with zero in either table condition. These Study II counts differ from the 16-case post-hoc intersection in Study I. Agreement therefore requires checking against the supplied evidence.

Discussion

Preservation and evidence export

Retrieving a value and establishing which record a saved citation denotes require different evidence. Study II supplies the ID-to-record relation without adding document or user text. For the tested questions, this restores a directly checkable link while leaving the underlying tool content and actions unchanged.

Evidence collection and export should retain the native tool calls and results, their execution order, the authentic request and channel identity, the public actor account, and execution-scoped input-ID bindings. Table 2 describes the binding relation implemented here. Generated accounts should remain identifiable as generated artifacts, and derived views should identify withheld fields. A locator alone does not supply the content of the record it names.

The experiment removes bindings from derived evidence packets rather than measuring their loss in deployed systems. Its preservation implication concerns retaining the relationships needed to check a finding as records are transformed or shared. Hashes and retained originals support alteration detection within the instrumented collection. Source authenticity and operational chain of custody require separate validation.

Reporting the supported relation

The illustrative case **rcase-001** separates three reportable questions (Table 8). Both packets establish that the recipient appears in the earlier bill result. Only the packet with the binding table establishes that the saved Basis refers to that record. The finding describes the recorded citation relationship. It does not establish why the actor selected the recipient.

Table 8: Findings supported by the C00/C10 packets for **rcase-001**. These summarize the existing illustrative case under the declared evidence contract.

Question	Supported finding
Q1: Request	Unresolved in both packets: authentic request text is withheld.
Q2: Occurrence	The recipient appears in act-01.result before the transfer in both packets.
Q3: Citation	Unresolved in C00; the explicit binding in C10 identifies the bill result as the saved Basis reference.

Reporting should also distinguish an empty relation, an observed absent Basis field, and insufficient supplied evidence. An empty set reports the result of the defined

record query. A missing field reports the state of the saved account. Neither warrants extending the finding to unobserved actor behavior.

Validation of assisted findings

Accurate literal retrieval, complete-reference agreement, and cross-reader agreement each coexist with unsupported citation findings. Evaluation should therefore report supported definite findings and justified abstention separately. Adding evidence can make a previously correct guess supportable without producing a comparable increase in complete-reference agreement.

For these literal and identifier questions, the packet-only deterministic comparator answered or abstained correctly throughout. It supplies a validation check for relationships that can be established through matching and table joins. The readers still produce 12–16 nongrounded Q3 answers with complete tables, so preserving the required relation and reading it correctly are separate requirements. Findings that depend on an unavailable binding should retain an unresolved status in the assisted output.

For reconstruction review, a generated finding can be retained together with the case and target-action identifiers, the supplied packet, and the locator set it asserts. Verification can then establish literal occurrence in earlier records before checking any execution-local citation binding. This makes a citation claim inspectable as a relation between preserved records. It also keeps the actor’s generated account distinct from the independent check of that account. Requirement-based testing of search functions [10] provides a precedent for validating these operations separately.

A review output should preserve the distinction between supported nonempty relations, supported empty relations, observed Basis absence, and insufficient supplied evidence. These outcomes have different consequences for reporting. An empty intersection answers the defined query, whereas an unresolved binding leaves a relation unestablished even when the corresponding source is plausible. Keeping the status beside the cited locator allows a later reviewer to see whether a definite finding was justified by the packet, rather than reconstructing that distinction from a fluent narrative.

Support should also be recorded against the evidence version used for the finding. In **rcase-001**, supplying the table changes the support for Sonnet’s unchanged answer. A later packet can therefore justify a relation that an earlier packet could not establish. Recording which content and bindings were supplied makes that change inspectable without treating the earlier guess as grounded. This documentation perspective is consistent with work on reliability validation of forensic interpretation [20]. It locates the reported finding within a reproducible evidence check.

One potential application is reviewing a reduced agent trace after an automated transfer. If the native execution retained the required citation bindings, an exporter can

include them with the records used to verify a finding. If a required binding was never recorded, this audit leaves the relation unresolved. This use depends on preserving execution-scoped bindings during collection, before producing an assisted report.

Limitations

The audit verifies literal occurrence, temporal eligibility, and recorded identifier relations. Its findings describe what the supplied packet supports. A saved citation records the actor’s stated reference and does not establish semantic authorization, harmful intent, or internal causal reliance.

The corpus contains 64 applicable cases from 13 tasks in one simulated Banking environment, with 32 structurally inapplicable inventory entries retained in the selection manifest. The derived evidence conditions do not estimate the prevalence of missing bindings in operational systems. No investigator evaluation was conducted, so the results do not quantify benefits to casework.

The studies were developed retrospectively, with inputs and analyses fixed before each pass. Sonnet was selected after Luna’s results. Repeated readings reuse the same cases, and task-bootstrap intervals are exploratory. The comparison describes two specific reader configurations, with Sonnet sharing the Claude family with the Haiku actor in 17 cases. One response per condition and hosted-model nondeterminism limit the attribution of individual output changes, particularly under identifier renaming.

Conclusion

Forensic review of a saved agent action requires distinguishing a value recorded in the evidence from a supported link to the actor’s stated citation. In this corpus, exact literal retrieval and agreement with the complete reference coexist with unsupported citation findings, including answers shared by both readers. In C00, Sonnet retrieves every literal-location set but makes 26 unsupported citation-source assertions in the 28 mapping-required cases. Twenty-two match the complete reference. This contrast shows why validation must retain the support status alongside the answer when checking whether the supplied records establish the reported relation. Explicit ID-to-record bindings make the relation checkable and improve grounded reconstruction for the measured readers, although complete tables do not eliminate their errors. Identifier renaming alone provides no consistent remedy.

For the tested questions, preservation should retain record content, execution order, channel identity, and the execution-scoped bindings needed to verify a reported relation. Generated accounts should remain distinguishable from the records they reference. The packet-only deterministic comparator provides a check for the literal and identifier relations examined here. Reporting supported definite findings, observed empty or missing states, and

unresolved relations separately prevents agreement with a complete reference from being treated as evidence of support in the supplied packet.

References

- [1] Abolghasemi, A., Azzopardi, L., Hashemi, S.H., de Rijke, M., Verberne, S., 2025. Evaluation of Attribution Bias in Generator-Aware Retrieval-Augmented Large Language Models, in: Findings of the Association for Computational Linguistics: ACL 2025, pp. 21105–21124.
- [2] Casey, E., Barnum, S., Griffith, R., Snyder, J., van Beek, H., Nelson, A.J., 2017. Advancing coordinated cyber-investigations and tool interoperability using a community developed specification language. *Digital Investigation* 22, 14–45. doi:10.1016/j.diin.2017.08.002.
- [3] Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S.R., Leike, J., Kaplan, J., Perez, E., 2025a. Reasoning Models Don’t Always Say What They Think. *arXiv:2505.05410*. preprint.
- [4] Chen, Y., Li, H., Li, Y., Liu, Y., Song, Y., Hooi, B., 2025b. TopicAttack: An Indirect Prompt Injection Attack via Topic Transition, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 7327–7345.
- [5] Debenedetti, E., Zhang, J., Balunovic, M., Beurer-Kellner, L., Fischer, M., Tramèr, F., 2024. Agent-Dojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents, in: Advances in Neural Information Processing Systems, pp. 82895–82920.
- [6] Field, C.A., Welsh, A.H., 2007. Bootstrapping Clustered Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 69, 369–390.
- [7] Gao, T., Yen, H., Yu, J., Chen, D., 2023. Enabling Large Language Models to Generate Text with Citations, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 6465–6488.
- [8] Goo, T., Kim, J., Han, K., Jo, G., Kim, J.S., Kim, T.H., 2026. Evaluating and Guarding Citation Faithfulness in Agentic Scientific Synthesis. *arXiv:2607.20527*. preprint.
- [9] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M., 2023. Not What You’ve Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection, in: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, pp. 79–90.
- [10] Guo, Y., Slay, J., Beckett, J., 2009. Validation and verification of computer forensic software tools—searching function. *Digital Investigation* 6, S12–S22. doi:10.1016/j.diin.2009.06.015.

- [11] Liao, J., 2026. Auditing Provenance Sensitivity in LLM Agent Action Selection. *arXiv:2607.20827*. preprint.
- [12] Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P., 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12, 157–173.
- [13] Ma, X., Li, T., Xiao, C., Yu, Z., Zhang, N., Vorobeychik, Y., 2026. AutoDojo: Adaptive Black-Box Attacks Reveal the Limits of IPI Defenses and Task-Specification Effects in LLM Agents. *arXiv:2606.15057*. preprint.
- [14] Miceli Barone, A.V., Barez, F., Cohen, S.B., Konstas, I., 2023. The Larger they are, the Harder they Fail: Language Models do not Recognize Identifier Swaps in Python, in: *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 272–292.
- [15] Michelet, G., Breiteringer, F., 2024. ChatGPT, Llama, can you write my report? An experiment on assisted digital forensics reports written using (local) large language models. *Forensic Science International: Digital Investigation* 48, 301683. doi:10.1016/j.fsidi.2023.301683.
- [16] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.t., Koh, P., Iyyer, M., Zettlemoyer, L., Hajishirzi, H., 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation, in: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100.
- [17] Mishra, A., Sharad, K., 2026. Observability for Delegated Execution in Agentic AI Systems. *arXiv:2606.09692*. preprint.
- [18] Moreau, L., Missier, P. (Eds.), 2013. PROV-DM: The PROV Data Model. World Wide Web Consortium. URL: <https://www.w3.org/TR/2013/REC-prov-dm-20130430/>. W3C Recommendation, 30 April 2013.
- [19] Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., Zhang, T., 2024. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models, in: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878.
- [20] Nordvik, R., Stoykova, R., Franke, K., Axelsson, S., Toolan, F., 2021. Reliability validation for file system interpretation. *Forensic Science International: Digital Investigation* 37, 301174. doi:10.1016/j.fsidi.2021.301174.
- [21] Panickssery, A., Bowman, S.R., Feng, S., 2024. LLM Evaluators Recognize and Favor Their Own Generations, in: *Advances in Neural Information Processing Systems*, pp. 68772–68802.
- [22] Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., Petrov, S., Tomar, G.S., Turc, I., Reitter, D., 2023. Measuring Attribution in Natural Language Generation Models. *Computational Linguistics* 49, 777–840.
- [23] Scanlon, M., Breiteringer, F., Hargreaves, C., Hilgert, J.N., Sheppard, J., 2023. ChatGPT for digital forensic investigation: The good, the bad, and the unknown. *Forensic Science International: Digital Investigation* 46, 301609.
- [24] Studiawan, H., Breiteringer, F., Scanlon, M., 2025. Towards a standardized methodology and dataset for evaluating LLM-based digital forensic timeline analysis. *Forensic Science International: Digital Investigation* 54, 301982.
- [25] Turpin, M., Michael, J., Perez, E., Bowman, S., 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting, in: *Advances in Neural Information Processing Systems*, pp. 74952–74965.
- [26] Wallat, J., Heuss, M., Rijke, M.d., Anand, A., 2025. Correctness is not Faithfulness in Retrieval Augmented Generation Attributions, in: *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pp. 22–32.
- [27] Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T., Sui, Z., 2024. Large Language Models are not Fair Evaluators, in: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450.
- [28] Wang, P., Li, Y., Tian, Y., 2026. Aligning Provenance with Authorization: A Dual-Graph Defense for LLM Agents. *arXiv:2605.26497*. preprint.
- [29] Zhan, Q., Liang, Z., Ying, Z., Kang, D., 2024. InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents, in: *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10471–10506.
- [30] Zhang, H., Huang, J., Mei, K., Yao, Y., Wang, Z., Zhan, C., Wang, H., Zhang, Y., 2025. Agent security bench ASB: Formalizing and benchmarking attacks and defenses in llm-based agents, in: *International Conference on Learning Representations*, pp. 35331–35366.