American Journal of Innovation in Science and Engineering (AJISE)

ISSN: 2158–7205 (ONLINE)

VOLUME 5 ISSUE 3 (2026)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Innovation in Science and Engineering (AJISE)

Volume 5 Issue 3, Year 2026
ISSN: 2158-7205 (Online)
DOI: https://doi.org/10.54536/ajise.v5i3.8542
https://journals.e-palli.com/home/index.php/ajise

Presentation Order and Structured Diagnosis in a Synthetic Payment Workflow Benchmark

Ayush Ojha[1*]

Article Information

Received: July 07, 2026

Accepted: September 08, 2026

Published: September 21, 2026

Keywords

Financial Technology, Local Language Model, Payment Workflow, Root-Cause Analysis, Synthetic Benchmark

ABSTRACT

We compare three orderings of identical synthetic payment-event records across 96 cases, four local model builds, two seeds, and 2,304 calls under a fixed 256-token output limit. On 80 labeled cases, trace ordering increased exact diagnosis from 187/640 (29.22%) to 248/640 (38.75%), a 9.53-percentage-point gain. Safe abstention on 16 underdetermined cases rose from 21/128 (16.41%) to 23/128 (17.97%), satisfying the registered five-percentage-point relative margin and numerical joint gate. However, three builds never safely abstained, and structurally valid commitments increased from 50/128 to 68/128. All 803 length-terminated outputs were invalid; 47 of the net 61 additional correct outputs crossed the validity boundary. Post-hoc structural diagnostics identified fixed templates, a visible marker separating every underdetermined case, and decoy timestamps that place all distractors last under trace ordering. The protocol's intended exhaustive two-alternative ambiguity construction was not established, although all 16 observations remained non-identifying. These results demonstrate end-to-end presentation sensitivity within the recorded suite and response budget. They do not isolate temporal reasoning, establish general ambiguity recognition, or demonstrate operational safety.

INTRODUCTION

Payment workflows distribute one customer-visible outcome across checkout, risk, routing, authorization, capture, callback, settlement, and ledger services. When an unexpected outcome occurs, relevant facts can be scattered across services or mixed with harmless activity. Operational root-cause analysis requires deciding which observations support a cause and whether the available record identifies it uniquely. A closed synthetic classification task can test parts of this interface without establishing the realism or difficulty of operational diagnosis.

Language models are increasingly studied as assistants or agents for root-cause analysis (RCA). The attraction is straightforward: operational evidence is heterogeneous and partly textual, while a model can consume codes, timestamps, attributes, and short policy descriptions in one context. Yet final-answer accuracy alone does not establish that a diagnosis is supported. A model may name the right component from a superficial cue, cite irrelevant evidence, or express certainty when the observed events do not distinguish two plausible causes. Recent work on cloud and microservice RCA has made this distinction visible by adding fault-propagation annotations, analyzing diagnostic trajectories, and separating data ambiguity from reasoning gaps (Fang *et al.*, 2026; Gopal & Krishnan, 2026; Lu *et al.*, 2026).

This study asks whether presentation order changes structured diagnosis on a fixed synthetic payment-event set. Each case is shown as a shuffled mixture with decoys, a chronological trace, or a deterministic service grouping. All arms contain identical event objects. The model must select a catalog condition, cite exact event identifiers, and abstain on incomplete observations that do not establish one condition. The catalog includes rejected attempts, intended control responses, and anomaly states; its fault labels do not uniformly denote malfunctioning components or failed settlement.

The design makes three deliberate reductions. First, it uses a synthetic generator rather than production logs. That choice gives exact intervention labels and prevents customer data, credentials, or institution-specific configuration from entering the study, but limits external validity. Second, it evaluates short, fixed-context diagnosis rather than an agent that queries tools. That removes search-policy and tool-use confounding, but does not measure full incident-response behavior. Third, it scores only machine-checkable fields. No endpoint relies on a language model judge, and the paper makes no free-form explanation-quality claim.

These reductions turn evidence presentation into an information-systems design variable rather than a proxy for an entire incident-response product. Operational platforms routinely transform the same underlying records into chronologies, service views, dashboards, and alert bundles. Such transformations can reduce an operator's sorting burden, but they can also relocate decisive events, concentrate superficially similar records, or make a late symptom visually prominent. If a language model is sensitive to those changes, an observed diagnosis can depend partly on the presentation layer even when the available evidence is unchanged. The paired design

[1] Georgia Institute of Technology, Atlanta, Georgia, United States
[*] Corresponding author's e-mail: aojha47@gatech.edu

therefore asks about a property of the combined evidence interface and model, not about the model in isolation.

Identifiability and response formatting are separate concerns. The incomplete cases warrant withholding a unique label, whereas labeled cases supply the attributes defining one catalog condition. However, the incomplete cases also expose a missing-data marker, and both subsets repeat narrow templates. The experiment therefore measures behavior under this specific evidence contract, rather than establishing that success requires general causal reasoning or realistic incident reconstruction.

This separation also makes the practical question more conservative. A chronological view would not be useful merely because it increased the number of fault labels emitted. Any gain in exact diagnosis must coexist with no material deterioration in the system's willingness to withhold a fault label when the visible record does not justify one. The registered joint gate encodes that requirement directly. It does not assert that the selected margin is a production safety threshold; it is a pre-outcome benchmark rule for this finite suite.

The study contributes a reproducible measurement of an evidence interface. It provides deterministic payment-event cases, paired presentations over identical objects, separate label and supporting-event endpoints, and a registered abstention rule. These endpoints distinguish selected fields of a response; neither a correct label nor a formally valid abstention guarantees a complete evidentiary justification.

The primary estimand is the exact finite-suite paired difference in fault identification between trace-ordered and raw-mixed evidence across 80 unambiguous frozen cases, four pre-outcome-locked local model builds, and two decoding seeds. Sixteen additional cases are deliberately underdetermined and support a safe-abstention noninferiority gate. The study reports malformed outputs and failed calls rather than removing them. service_grouped is secondary. Neither cases nor model builds are treated as population samples.

The contribution is a payment-specific, protocol-bound comparison of presentation order over identical event objects, with exact fault and event-ID scoring and a separate missing-discriminator abstention criterion. Controlled RCA, financial evidence contracts, fault injection, and order sensitivity are established ideas. The closest-work comparison below locates the contribution at their joint experimental design without asserting global novelty.

LITERATURE REVIEW

Outcome labels and causal evidence in root-cause analysis

RCA benchmarks have often emphasized whether a method names the correct service or fault. That endpoint is necessary, but it can conceal the route by which a symptom arose. OpenRCA 2.0 addresses this problem with PAVE, a stepwise labeling protocol that reconstructs causal propagation paths from known interventions (Fang *et al.*, 2026). Its authors report a substantial gap between recovering at least one correct root-cause service and grounding that service in a verified path. The implication for the present study is not that one benchmark transfers directly to payments. Rather, it is that answer identity and evidentiary support should be separate endpoints.

Lu *et al.* (2026) study 3,500 agent trajectories and likewise argue that endpoint correctness can hide poor diagnostic quality. Their trajectory-level framework compares investigation behavior with manually curated fault-propagation paths, and their DiagGuard architecture separates grounding from localization and verification. The present experiment removes interactive investigation and tests a more basic presentation layer. Its evidence precision and recall measure overlap with generator-assigned reference events, rather than validating a fault-propagation path or an internal reasoning trace.

TraceBench provides another controlled perspective (Bendinelli *et al.*, 2026). It generates root-cause tasks from interpretable physical dynamical systems and examines how LLM agents analyze time-series observations. TraceBench and this study share the value of known interventions and controlled conditions, but differ in domain, input structure, and endpoint. This study uses payment-event traces rather than physical time series, keeps all three presentations paired to one incident, and requires exact event identifiers and explicit abstention.

FinRCA-Bench is the closest retrieved financial-operations neighbor (Ghawate, 2026). It provides 2,250 deterministic synthetic reconciliation cases spanning accounts-payable and bank records, including injected causal failures, legitimate cases, and hard negatives. Its evaluation separates evidence retrieval from reasoning and includes exact class and record-level evidence contracts. That work establishes that synthetic financial RCA, deterministic failure injection, negative cases, and auditable evidence contracts are not new in isolation. Its controlled model comparison changes which records are retrieved by dense and typed relational retrieval methods. The present study instead supplies the identical complete event objects in each arm and changes only their order. It therefore does not compare retrieval architectures and cannot claim to introduce financial RCA or evidence contracts.

Financial-document retrieval provides a related evaluation setting. Mohammed and Lund (2026) assess a chatbot for compensation queries using separate RAGAS measures of context recall, faithfulness, context utilization, and context precision. The present benchmark instead supplies all event objects directly and scores exact condition labels and event identifiers, so it does not evaluate retrieval quality or the same response-quality measures.

Model and scaffold failure modes

The performance of an RCA system reflects more than its language model. Kim *et al.* (2026) analyze 1,675 OpenRCA agent runs and identify recurring failures across intra-agent reasoning, inter-agent communication, and agent–

environment interaction. They report that hallucinated data interpretation and incomplete exploration persist across model tiers, while communication changes help only part of the problem. Riddell *et al.* (2026) similarly compare agentic and non-agentic workflows over a large controlled scenario set and develop a taxonomy of reasoning failures. These studies motivate a design that holds the scaffold deliberately small. By presenting all evidence in one request, the current benchmark cannot attribute a failure to query planning or tool selection. Its claims are correspondingly limited to fixed-context diagnostic behavior.

Gopal and Krishnan (2026) distinguish a reasoning gap, where relevant evidence is available but unused, from data ambiguity, where the evidence does not support a unique answer. That distinction motivates the two subsets here. Eighty labeled cases contain catalog-specific conditions and reference events; 16 incomplete observations contain no injected true fault. Their registered target is abstention. Because a visible missingness marker perfectly identifies the latter subset, this study cannot separate general ambiguity recognition from response to an explicit cue.

Safe-Psych provides a current benchmark precedent for separating diagnosis from requests for more evidence or abstention (Presacan *et al.*, 2026). It segments more than 1,000 psychiatric notes to model sequential evidence disclosure and assigns DIAGNOSE, CLARIFY, or ABSTAIN decisions at successive stages. Its setting, data source, and clinical risk are different from the synthetic payment suite, and the present model cannot ask a follow-up question. The relevant common point is narrower: a diagnostic benchmark can explicitly evaluate whether the available information warrants commitment. The current study operationalizes that question with deterministic missing-discriminator cases and a binary safe-abstention contract rather than clinician-derived sequential action labels.

Gong *et al.* (2026) provide a close operational-telemetry comparison. ORCA-bench evaluates agents that query metrics, logs, traces, and source code from an OpenTelemetry-instrumented microservice system. Its ground-truth construction includes model-generated material: expert SREs validate symptom rubrics, and a 40-task Verified subset receives full human confirmation. The present study does not introduce operational-telemetry RCA benchmarking. Its distinction is a payment-specific, non-agentic, same-event order intervention with exact fault and event-ID scoring and explicit missing-discriminator cases.

Evidence organization and distributed traces

Distributed tracing supplies a common language for correlating operations. OpenTelemetry describes a trace as the path of a request through an application, assembled from spans carrying trace and parent identifiers, timestamps, attributes, events, links, and status (OpenTelemetry, 2026). The synthetic records in this study borrow the general notions of correlation, timestamps, services, and event attributes. They are not OTLP records and are not presented as a conformance test.

Evidence presentation changes the arrangement a model receives. A shuffled view interleaves records, a chronological view supplies temporal order, and service grouping clusters records by component. Every arm contains the same event objects. These are order-only interventions, including the secondary grouping arm, rather than competing RCA products. Chronological ordering also segregates this generator's late-timestamped decoys, so it does not isolate the benefit of reconstructing a timeline.

TELLER demonstrates that temporal and structural trace representation is already an active RCA technique (Xu *et al.*, 2026). It reconstructs cross-layer call-chain trees for language-model inference services and forms causal-context slices that preserve parent-child structure, temporal order, and communication relations. Its representation is designed to localize faults across execution layers; it does not present an identical complete event set in paired orders. This difference matters for attribution. The present study tests whether an order-only interface change affects a fixed diagnosis task, whereas TELLER evaluates a purpose-built structural representation within a broader RCA system.

Liu *et al.* (2024) show that long-context question-answering performance can vary with the position of relevant information. Because changing event order also changes each event's prompt position, an observed raw-mixed versus trace-ordered difference may reflect position sensitivity as well as the burden of sorting events. Order effects do not by themselves prove temporal or causal understanding. The study therefore estimates a presentation-order effect on its finite suite and does not identify a cognitive mechanism.

Han *et al.* (2026) make the order issue more direct. Their biomedical study presents the same correct and contradictory documents in opposite orders and finds that predictions can flip even though the evidence set is unchanged. They also evaluate conflict-aware selective abstention. This prior work means that identical-evidence order sensitivity and abstention under difficult evidence are not individually novel. Its intervention, however, concerns two mutually conflicting biomedical documents. It does not evaluate a multi-event payment workflow, exact supporting-event recovery, or a coherent missing-discriminator case in which two diagnoses remain compatible because a required fact is absent. Together, Han *et al.* (2026) and Liu *et al.* (2024) require the present paper to describe its contrast as presentation-order sensitivity, not evidence that a model learned temporal causality.

Payment-specific diagnosis and scope

The payment catalog contains eight synthetic conditions: expired authorization, amount-scale mismatch, a rejected duplicate retry, an active risk hold, callback-integrity

failure, ledger posting delay, routing-policy denial, and issuer decline. These are benchmark categories with machine-checkable observations. A rejected retry can represent successful duplicate suppression, and a risk hold or routing denial can be an intended safeguard. The catalog does not establish that an observed control decision was erroneous or that the lower-level cause of a symptom has been identified.

Condition identity and evidence identity remain different endpoints. The generator records event identifiers that instantiate each condition and keeps this reference set outside the model-visible payload. Evidence scoring asks whether the response points to those records. It does not establish a uniquely minimal explanation: checkout can supply currency context even when authorization and capture already establish an amount discrepancy. Callback-integrity and duplicate-retry cases also retain successful settlement and ledger posting, so their labels must not be interpreted as proof of overall payment failure.

The benchmark deliberately avoids several adjacent payment questions. It does not estimate fraud, creditworthiness, settlement risk, chargeback probability, or customer harm. It does not reconstruct a real processor's internal state, compare providers, or recommend an operational action. FinRCA-Bench already covers a broader financial-reconciliation setting in which retrieval architecture determines which records reach the reasoning model (Ghawate, 2026). The current study's narrower payment contribution is the paired ordering experiment over a fixed complete record, not financial-domain coverage by itself.

The target also differs from fraud screening. Serifat *et al.* (2025) review supervised, unsupervised, reinforcement-learning, and ensemble approaches to detecting fraud in digital banking. Here, the labels identify synthetic workflow conditions, including legitimate control responses, rather than whether a transaction is fraudulent. The study does not claim that this catalog is complete, institutionally representative, or validated by a payment-operations panel. It also does not use chargeback records, cardholder data, bank data, processor credentials, or actual provider codes. Class names and codes are benchmark constructs. These boundaries separate the study from fraud detection, credit decisions, compliance testing, and live incident-response evaluation.

Closest-work conclusion

The pre-outcome search identified close work on controlled attribution, causal path annotation, trajectory-level diagnosis, production-fidelity operational telemetry, financial-reconciliation RCA, ambiguity, identical-evidence order effects, and incomplete-evidence abstention. FinRCA-Bench is the closest retrieved financial-domain prior, ORCA-bench is the closest retrieved operational-telemetry prior, and "When Evidence Conflicts" is the closest retrieved order-plus-abstention prior. Each materially narrows the contribution.

The inspected primary sources do not jointly evaluate these four elements: payment-workflow faults; paired orderings of identical complete event objects; exact fault and supporting-event-ID scoring; and deliberately underdetermined cases with a prespecified safe-abstention gate. This paper evaluates that bounded combination. It does not introduce financial RCA, operational-telemetry benchmarking, evidence contracts, fault injection, order sensitivity, temporal trace representation, or abstention. Its contrast measures end-to-end presentation sensitivity under a fixed response contract and output budget, rather than an isolated improvement in temporal or causal reasoning.

MATERIALS AND METHODS

Protocol status and research question

Protocol v1.0.2 was invalidated before aggregate inspection because of design and runner defects. Its frozen tree and 430-row journal remain byte-preserved and are excluded from analysis and claim tuning. Corrected v2.0.0 and v2.0.1 were superseded before any corrected-study call to strengthen metadata validation, integrity exports, authorization, and raw-JSON preservation; both remain outcome-ineligible. Authoritative v2.0.2 binds the protocol, schema, generator, runner, scorer, integrity validator, analysis, 96-case casebook, 2,304-job plan, and runtime/model lock under artifacts/frozen-v2.0.2/ (manifest SHA-25638f2fac355e9519f876d3e4eca55ca5f15a7ce66028dc88a719cd7d206c11b84). Its amendment and detached anchor document zero corrected-study outcomes before freeze eligibility. Mismatches stop execution; design changes require a new version preserving prior lineages. Analysis uses only the complete, authenticated v2.0.2 journal.

Registered means specified in the internally frozen protocol, not a public registry deposit, third-party timestamp, or digital signature. The study tests whether trace ordering improves exact diagnosis without materially reducing safe abstention. Construction and interpretation limitations identified post hoc are reported separately.

One joint confirmatory comparison is registered: trace_ordered - raw_mixed. It passes only if both conditions hold:

1. The exact finite-suite accuracy difference on 80 unambiguous cases is greater than zero; and
2. The exact finite-suite safe-abstention difference on 16 underdetermined cases is at least -0.05.

The five-percentage-point safety margin is a benchmark choice, not an operational safety threshold. service_grouped is secondary. Null, negative, and safety-gate-failing results remain reportable.

For each case-model-seed combination, the paired requests differ only in event order, holding event membership, catalog, schema, decoding settings, and identifiers fixed. The intervention is the case presentation. It removes between-case difficulty and retrieval differences from the contrast but jointly changes token position, adjacency, and encounter order; the estimand does not isolate a

reasoning mechanism.

Case generation

A fixed-seed deterministic generator creates 96 cases. Eighty comprise ten instances of eight templates: authorization expiry, currency-scale inconsistency, duplicate retry, risk hold, callback-integrity failure, ledger posting delay, merchant routing, and issuer decline. Dates, identifiers, amounts, and some region and decoy attributes vary; mechanisms, fault magnitudes, thresholds, USD currency, and merchant category do not. There are no fault-free controls or structurally novel mechanisms.

For each labeled case, the generator modifies baseline workflow attributes or timing and stores reference event identifiers outside the visible payload. Blocking conditions remove incompatible downstream stages. A rejected repeat capture retains the accepted original capture and settlement; callback-verification rejection retains settlement and posting, without specifying an independent source of ledger success information. Direct labels and gold fields are hidden, but literal codes, statuses, and class-specific stages remain informative.

Sixteen observations comprise four variants of four incomplete templates: generic gateway failure, capture rejection, policy block, and incomplete reconciliation. Hidden metadata names two intended alternatives and a missing discriminator; neither alternative is injected as true, fault_id is null, and gold evidence is empty. No complete latent fault realization is retained for reproducing the observation by masking.

Visible event, trace, idempotency, and settlement identifiers are deterministic SHA-256-derived opaque tokens in a v2 namespace, without case indices, fault families, ambiguity groups, or gold markers.

The incomplete observations do not identify a unique cause. Naming two alternatives does not exclude other catalog causes, so the intended exactly-two-alternative construction was not established. The original abstention labels and numerical gate are retained for these non-identifying observations without retrospectively validating or amending the construction.

Every incomplete case, and no labeled case, contains missing_fields=true: a perfect visible abstention-subset cue. The study therefore does not isolate general identifiability reasoning, nor establish whether a model used the cue.

Automated invariants check equal event-object sets, opaque identifiers, absence of direct label/gold fields, class/group balance, and stage constraints. They do not establish cue-free construction, exhaustive causal compatibility, or operational realism. Post-hoc structural diagnostics leave cases and scoring unchanged.

Evidence arms

Every case produces three views:

- Raw mixed: all case and decoy events in deterministic shuffled order.
- Trace ordered: the same full event set sorted by timestamp and event identifier.
- Service grouped: the same full event set sorted by service, timestamp, and event identifier.

Arms contain byte-equivalent event objects, with no label- or gold-based selection. Raw shuffling is seeded, not randomized anew at inference; the listed sort keys provide deterministic tie-breaking. Only raw mixed versus trace ordered is confirmatory.

All four decoys are timestamped after the last substantive event, producing a trailing block in 96/96 trace-ordered cases versus 1/96 raw-mixed cases. Reduced distractor interleaving may contribute to the contrast; isolating it from chronology requires a separate intervention.

Model suite and inference

Four local chat-model builds and two decoding seeds were fixed before outcomes, with no favorable model subset selected. Model names enter through PAYDIAG_MODELS, with no application default. A frozen inventory-only lock records their unique names, digests, metadata, Ollama and Python versions, platform, endpoint origin, and timestamp. Execution requires locked Python 3.14 and exact model name, order, digest, Ollama-version, and endpoint matches.

The factorial contains 96 cases × 3 arms × 4 builds × 2 seeds = 2,304 calls. Each request uses temperature 0.2, a 256-token output limit, the common fault catalog, the JSON schema as Ollama format, and think=false. The scorer independently validates returned text against exactly four fields: fault_id, evidence_ids, abstain, and confidence. Provider schema constraints do not guarantee valid or complete output.

Cases are deterministically shuffled within model blocks; arms are exactly counterbalanced across all six within-case arm-by-seed positions. Each v2 job identifier hashes the order seed, case, arm, model, and decoding seed.

The runner imports verified frozen source. A non-dry run requires both --execute and the frozen PAYDIAG_EXECUTION_ACK before freeze verification, inventory access, journaling, or requests; --dry-run and --execute are mutually exclusive. An atomic exclusive lock permits one journal writer. Raw provider message strings are preserved, and duplicate JSON keys are rejected. Resume requires exact frozen job tuples and run metadata; duplicates, unknown rows, missing metadata, or tuple drift stop execution. Failures, timeouts, malformed JSON, duplicate keys, and schema failures remain terminal incorrect and invalid responses, without silent retries under the same identity.

Requests provide the catalog, one event view, and the four-field schema, without chain-of-thought, tools, retrieval, or remediation. Confidence is validated and retained in the authenticated raw envelope but not analyzed: one scalar cannot establish calibration.

Endpoints

Primary exact fault_id accuracy uses 80 labeled cases. Correct responses must be structurally valid, coherent,

non-abstaining, and match the injected label. Coherence requires visible cited identifiers and consistent abstention/label fields. Evidence overlap is scored separately; incomplete or empty evidence can accompany exact diagnosis.

On an underdetermined case, the response categories are explicit and disjoint under the valid-response contract:

- **safe abstention:** structurally valid, abstain=true, and fault_id=null;
- **explicit commitment:** structurally valid and either abstain=false or fault_id is non-null, including contradictory combinations; or
- **invalid response:** failure of the exact structural contract, counted in neither valid-response category.

Secondary endpoints cover service-grouped contrasts, commitment and invalid-response rates, labeled-case evidence precision/recall, validity/coherence, and model/arm failures. Duplicate JSON keys are rejected; duplicate or non-visible evidence IDs violate validity or coherence under the frozen contract. Underdetermined cases have no gold evidence and receive safety-category rather than causal-event recall scores.

Evidence precision is returned-ID/gold-set intersection divided by all returned distinct IDs, including non-visible IDs; recall divides that intersection by gold-set size. Empty sets and structurally invalid responses score zero on labeled cases. Gold events are generator references, not uniquely minimal explanations. For amount mismatch, authorization and capture establish the discrepancy while checkout adds context; citing only the numeric records gives recall 2/3. Label accuracy and event recovery can diverge.

Validity checks exact JSON structure, keys, types, values, and uniqueness. Safe abstention requires only the valid decision fields above, without relevant evidence or a separate coherence requirement. Explicit commitment includes valid contradictory abstention/label combinations. Invalid truncated text may assert a fault but is excluded from that category. These definitions do not comprehensively measure caution or justification.

Statistical analysis

For each labeled case and arm, accuracy is averaged over eight model-seed responses; the primary estimate averages 80 paired trace_ordered - raw_mixed case differences. Safety uses the same calculation over 16 underdetermined cases. Cases receive equal weight; denominators are 640 labeled and 128 underdetermined responses per arm. The joint gate requires accuracy difference > 0 and safe-abstention difference ≥ -0.05.

Registered sensitivity analyses leave out one fault family for accuracy or one ambiguity group for safety. These diagnose subgroup dependence without additional hypothesis tests or claim gates. Service-grouped contrasts and arm-by-model summaries are secondary.

Cases and model builds form complete frozen target suites, not population samples; models and seeds are repeated measurements. No confirmatory bootstrap, p-value, or confidence interval is used. The exact finite-suite summaries support no population inference, and two seeds do not establish broad stochastic stability.

Missing or invalid terminal responses remain incorrect and invalid in their denominators, without answer imputation or silent retries. Coverage, provider completion, and response validity are distinct. Eligibility requires journal result.ok completion of at least 95% of calls for at least three of four builds. Provider-completed outputs may still be structurally invalid; neither invalid nor provider-failed rows are dropped.

Post-hoc diagnostics summarize done_reason and eval_count against validity and count raw-to-trace transitions among invalid, valid-incorrect, and correct states for all 640 labeled case-model-seed pairs. They retain registered rows and denominators and add no test or gate change. Conditioning on validity in both arms selects on a presentation-affected outcome and cannot identify a causal reasoning effect.

Reproducibility and validation

The implementation uses the Python standard library. Thirty unit tests cover generation, balance, opaque IDs, answer-field exclusion, incomplete construction, stage constraints, event equality, freeze integrity, locked env-driven models, the counterbalanced factorial, exclusive writing, frozen-source execution, authorization, dry-run/execution exclusion, raw-message preservation, duplicate-key rejection, response categories, journal validation, and registered analyses. Passing these checks does not validate exhaustive ambiguity construction or scientific claims. Frozen sources, protocol, casebook, and runtime lock are archived in the manifest.

Validation requires exact journal order/count, unique job IDs, case-arm-model-seed-position tuples, provider-model identity, completion eligibility, and provenance hashes; the analysis manifest records these bindings. Supplement S1 supplies cases, recorded responses, frozen scoring/analysis, post-hoc code, expected outputs, and offline replay instructions, without weights. Anonymous metadata removal and provenance rebinding are documented; numerical inputs and scoring remain unchanged, with original provenance retained privately.

Source reproduction regenerates the frozen casebook and job plan; outcome replay authenticates recorded responses and regenerates scores and summaries. Original analyses reproduced byte for byte in separate processes. The extracted anonymous package was checked in isolation on the same Windows computer with Python 3.12.14; inference used Python 3.14.5. This was local recorded-data replay, not fresh inference, a second implementation, or independent human replication. Supplement hashes distinguish original and anonymous provenance.

After extracting S1, follow README.md and run python -B reproduce.py --output ../reproduced-results. Replay verifies the input manifest, regenerates scores, finite-suite summaries and post-hoc diagnostics, and compares expected outputs using only the standard library, without

network or model service access. S2 independently replays structural cue/decoy counts from the unchanged casebook. S3 documents the deterministic 96-response sample, selection rule, unchanged outputs, frozen scores, reference overlap, and literal-ID checks; its README uses extracted S1 inputs. Neither performs inference or evaluates explanation relevance.

Ethics, privacy, and safety

The synthetic benchmark involves no human participants, animals, personal/customer data, real accounts or transactions, or credentials. Consent is inapplicable; no institutional approval or exemption is claimed. Models cannot use tools or execute actions, and the study makes no financial decision or recommendation.

Synthetic construction limits privacy exposure but does not establish operational harmlessness. The closed catalog does not test unseen or multiple faults, adversarial logs, provider-specific semantics, topology changes, or operator actions. Results establish neither production reliability nor customer impact, compliance, fairness, or safe automated remediation.

S1–S3 supply synthetic records, responses, code, and replay instructions for confidential review, without weights, third-party runtime binaries, or customer data. A private code repository is identified to the editor on the title page; anonymous supplements permit analysis replay without repository access.

Reporting and interpretation limits

Reporting follows the registered sequence: journal authentication and matrix coverage, paired accuracy and safety differences, then component and joint decisions. Both gates must pass; positive accuracy with failed safety is a joint failure. Secondary and post-hoc summaries do not change that decision, and selected response examples do not establish aggregates.

Sorting burden, event position, distractor adjacency, output completion, and model-specific prompt interactions may contribute; the design does not distinguish these mechanisms. Literal cues, repeated templates, and explicit missingness limit reasoning claims. Findings remain bounded by recorded cases, prompts, builds, runtime, and seeds.

Passed numerical gates do not certify ambiguity construction or scientific validity. The construction discrepancy remains reported, without claims of autonomous incident response, production readiness, internal causal reasoning, or universal trace-order benefit.

RESULTS AND DISCUSSION

All 2,304 recorded calls passed the exact matrix and provenance checks. Separate executions of the frozen analysis reproduced its outputs byte for byte. The results below preserve the registered endpoints and denominators; the additional termination and validity-transition analyses describe a reporting limitation discovered after outcome inspection.

Table 1 summarizes the study configuration and principal provenance bindings; complete manifests and replay instructions accompany Supplement S1.

Table 1: Study configuration and original provenance bindings.

Binding	Authenticated value
Analysis schema	paydiag-analysis-v2
Protocol version	2.0.2
Journal rows	2304
Scored rows	2304
Journal SHA-256	31a1f46fda63e874979f11c7cb72c14c404351063ce5bab8761c6ebb7d147b2d
Frozen manifest SHA-256	38f2fac355e9519f876d3e4eca55ca5f15a7ce66028dc88a719cd7d206c11b84
Ollama version	0.32.14
Eligible models	4
Confirmatory inference	exact finite-suite joint gate; no case bootstrap, p-value, or population interval
Cases	96
Models	4
Arms	3
Provider eligibility threshold	0.950000

Table 2 reports the registered confirmatory finite-suite decision.
Accuracy difference (trace_ordered - raw_mixed): 0.0953125.
Accuracy superiority gate: PASS.
Safe-abstention difference (trace_ordered - raw_mixed): 0.015625.
Registered safe-abstention margin: -0.050000.
Registered relative safe-abstention gate: PASS.
Joint confirmatory gate: PASS.

Table 2: Registered numerical decision on the recorded suite. PASS does not validate the intended exhaustive two-alternative construction or operational safety.

Field	Authenticated value
Estimand scope	exact finite frozen suite; no population inference
Confirmatory contrast	trace_ordered minus raw_mixed
Accuracy cases	80
Safety cases	16
Accuracy difference	0.0953125
Safety difference	0.015625
Safety margin	-0.050000
Accuracy superiority gate	PASS
Safety noninferiority gate	PASS
Joint gate	PASS

Table 3 reports model-wide coverage and response status, unambiguous exact-diagnosis and evidence endpoints, and underdetermined abstention endpoints using their prespecified case slices.

Table 3: Arm-by-model results and provider completion.

Descriptor	qwen3:8b	gemma3:4b	llama3.1:8b	phi4-mini:3.8b
Raw mixed / All cases / Rows	192	192	192	192
Raw mixed / All cases / Valid response	0.843750	0.062500	0.500000	0.885417
Raw mixed / All cases / Response coherence	0.843750	0.036458	0.484375	0.442708
Raw mixed / All cases / Invalid response	0.156250	0.937500	0.500000	0.114583
Raw mixed / Unambiguous cases / Rows	160	160	160	160
Raw mixed / Unambiguous cases / Exact fault accuracy	0.562500	0.018750	0.250000	0.337500
Raw mixed / Unambiguous cases / Evidence precision	0.599752	0.020867	0.295521	0.773542
Raw mixed / Unambiguous cases / Evidence recall	0.771875	0.068750	0.367708	0.646875
Raw mixed / Underdetermined cases / Rows	32	32	32	32
Raw mixed / Underdetermined cases / Safe abstention	0.656250	0.000000	0.000000	0.000000
Raw mixed / Underdetermined cases / Explicit commitment	0.281250	0.031250	0.531250	0.718750
Raw mixed / Underdetermined cases / Invalid response	0.062500	0.968750	0.468750	0.281250
Trace ordered / All cases / Rows	192	192	192	192
Trace ordered / All cases / Valid response	0.880208	0.213542	0.609375	0.963542
Trace ordered / All cases / Response coherence	0.880208	0.187500	0.609375	0.609375
Trace ordered / All cases / Invalid response	0.119792	0.786458	0.390625	0.036458
Trace ordered / Unambiguous cases / Rows	160	160	160	160
Trace ordered / Unambiguous cases / Exact fault accuracy	0.593750	0.156250	0.337500	0.462500
Trace ordered / Unambiguous cases / Evidence precision	0.601931	0.039918	0.345208	0.903542
Trace ordered / Unambiguous cases / Evidence recall	0.767708	0.196875	0.390625	0.759375
Trace ordered / Underdetermined cases / Rows	32	32	32	32
Trace ordered / Underdetermined cases / Safe abstention	0.718750	0.000000	0.000000	0.000000
Trace ordered / Underdetermined cases / Explicit commitment	0.281250	0.218750	0.687500	0.937500
Trace ordered / Underdetermined cases / Invalid response	0.000000	0.781250	0.312500	0.062500
Service grouped / All cases / Rows	192	192	192	192
Service grouped / All cases / Valid response	0.817708	0.119792	0.557292	0.906250
Service grouped / All cases / Response coherence	0.817708	0.093750	0.526042	0.520833

Service grouped / All cases / Invalid response	0.182292	0.880208	0.442708	0.093750
Service grouped / Unambiguous cases / Rows	160	160	160	160
Service grouped / Unambiguous cases / Exact fault accuracy	0.637500	0.087500	0.287500	0.318750
Service grouped / Unambiguous cases / Evidence precision	0.555551	0.019068	0.336778	0.835379
Service grouped / Unambiguous cases / Evidence recall	0.772917	0.125000	0.415625	0.696875
Service grouped / Underdetermined cases / Rows	32	32	32	32
Service grouped / Underdetermined cases / Safe abstention	0.718750	0.000000	0.000000	0.000000
Service grouped / Underdetermined cases / Explicit commitment	0.250000	0.093750	0.218750	0.656250
Service grouped / Underdetermined cases / Invalid response	0.031250	0.906250	0.781250	0.343750
Coverage / Matrix rows	576	576	576	576
Coverage / Expected rows	576	576	576	576
Coverage / Matrix coverage rate	1.000000	1.000000	1.000000	1.000000
Coverage / Provider success rate	1.000000	1.000000	1.000000	1.000000
Coverage / Valid response rate	0.847222	0.131944	0.555556	0.918403
Coverage / Invalid response rate	0.152778	0.868056	0.444444	0.081597
Coverage / Provider eligibility	PASS	PASS	PASS	PASS

All-case rows use 192 calls per arm-model cell and are shown only for response status; unambiguous rows use 160 calls per cell for exact diagnosis and evidence metrics; underdetermined rows use 32 calls per cell for safe abstention, explicit commitment, and invalid-response rates. Invalid responses remain in the relevant denominators and are not counted as abstentions. Evidence recall is therefore reported only for unambiguous cases with nonempty causal-event gold sets.

Table 3 shows substantial model differences under the same output contract. Provider completion was 100% for every build, but valid-response rates ranged from 0.131944 to 0.918403. Only qwen3:8b produced any safe abstention; the other three builds recorded none across all arms. The full table retains the exact diagnosis, evidence, validity, and abstention denominators rather than treating provider completion as successful diagnosis.

The registered joint decision is a relative finite-suite criterion, not a threshold for adequate absolute diagnosis or safe abstention. The full model-level results must be read alongside that decision; invalid responses and explicit commitments on underdetermined cases are distinct failure modes, neither of which is justified abstention.

Across the primary arms, the exact-accuracy difference is (248 - 187) / 640 = 0.0953125. The safe-abstention contrast is (23 - 21) / 128 = 0.015625. Across all arms, all 67 safe abstentions came from qwen3:8b; each other build recorded 0/96. All 67 also happened to satisfy the separate coherence check, although relevant evidentiary support is not guaranteed by either field-based measure. Structurally valid explicit commitments increased from 50/128 to 68/128 as invalid responses fell. This is not a count of every textual commitment, and the relative PASS does not establish adequate absolute safety.

Table 4 reports prespecified sensitivity and secondary contrasts.

Table 4: Sensitivity and secondary contrasts.

Analysis	Omitted group / contrast	Difference
Fault-family leave-one-out accuracy	authorization expired	0.098214
Fault-family leave-one-out accuracy	currency scale mismatch	0.094643
Fault-family leave-one-out accuracy	duplicate retry	0.080357
Fault-family leave-one-out accuracy	fraud hold	0.087500
Fault-family leave-one-out accuracy	issuer decline	0.101786
Fault-family leave-one-out accuracy	ledger lag	0.073214
Fault-family leave-one-out accuracy	merchant routing	0.085714
Fault-family leave-one-out accuracy	webhook integrity	0.141071
Ambiguity-group leave-one-out safety	capture rejection missing amount and retry context	0.020833
Ambiguity-group leave-one-out safety	gateway failure missing stage and code	0.031250
Ambiguity-group leave-one-out safety	policy block missing decision source	0.000000
Ambiguity-group leave-one-out safety	reconciliation gap missing integrity and latency	0.010417
Secondary service-grouped contrast	Accuracy: service grouped minus raw mixed	0.040625

Secondary service-grouped contrast	Accuracy: service grouped minus trace ordered	-0.054688
Secondary service-grouped contrast	Safety: service grouped minus raw mixed	0.015625
Secondary service-grouped contrast	Safety: service grouped minus trace ordered	0.000000

The secondary service-grouped arm exceeded raw mixed in accuracy by 0.040625 but was 0.0546875 below trace ordered, so grouping did not monotonically improve accuracy. One ambiguity-group leave-one-out safety contrast was zero. Neither secondary result alters the registered raw-versus-trace decision.

Post-hoc output-limit and response-validity diagnostics

Table 5 reports length termination by model and arm. Overall, 803/2,304 responses (34.85%) ended with provider done_reason=length and eval_count=256. Every such response was invalid under the frozen scorer, accounting for 803/891 invalid responses (90.12%). Length counts were 293/768 for raw mixed, 230/768 for trace ordered, and 280/768 for service grouped. The remaining 88 invalid responses had other termination outcomes. Length termination and invalidity co-occur here; these observations do not establish what each response would have become with a larger budget.

Table 5: Post-hoc length termination and invalid responses by model and arm.

Model build	Presentation	Calls	Length	Invalid
qwen3:8b	Raw mixed	192	30	30
qwen3:8b	Trace ordered	192	23	23
qwen3:8b	Service grouped	192	33	35
gemma3:4b	Raw mixed	192	160	180
gemma3:4b	Trace ordered	192	138	151
gemma3:4b	Service grouped	192	149	169
llama3.1:8b	Raw mixed	192	85	96
llama3.1:8b	Trace ordered	192	64	75
llama3.1:8b	Service grouped	192	81	85
phi4-mini:3.8b	Raw mixed	192	18	22
phi4-mini:3.8b	Trace ordered	192	5	7
phi4-mini:3.8b	Service grouped	192	17	18
All builds	All arms	2304	803	891

On unambiguous cases, valid responses increased from 369/640 under raw mixed to 421/640 under trace ordering. Table 6 retains all 640 pairs. Invalid-to-correct transitions numbered 60 and correct-to-invalid transitions 13, a net gain of 47. Valid-incorrect-to-correct transitions numbered 40 and the reverse 26, a net gain of 14. Together these account arithmetically for the original 61 additional correct outputs. Among the 327 pairs valid in both arms, raw and trace correct counts were 174 and 188.

Table 6: Post-hoc paired response states on 640 unambiguous case-model-seed pairs.

Raw state / trace state	Invalid	Valid incorrect	Correct	Total
Invalid	177	34	60	271
Valid incorrect	29	113	40	182
Correct	13	26	148	187
Total	219	173	248	640

The validity-conditioned subset is descriptive and does not replace the registered denominator. Because validity is itself affected by presentation, these transitions are not a causal decomposition of reasoning and formatting. The supported engineering result concerns the complete interface, model, response contract, and 256-token budget. Separately designed comparisons of output budgets and response controls would be required to isolate completion behavior from a reasoning claim; schema-constrained requests were already used here.

Post-hoc structural and response diagnostics

A post-hoc check constructed a simple lookup using missing_fields, literal gateway codes, risk hold, routing denial, callback rejection, and pending ledger status. It matched all 96 label-or-abstention annotations without timestamps, event order, opaque identifiers, or hidden gold fields as prediction inputs. The rule was chosen after inspecting the suite and emits neither evidence identifiers nor the required JSON response. It is a shortcut-existence witness, not a trained baseline, a prospective performance estimate, or a perfect exact-diagnosis score. Supplement S2 provides the script, unchanged casebook, and derived marker, cue, and decoy counts.

A separate post-hoc computational diagnostic examined 96 recorded outputs: three presentations for each of 32 cases comprising all 16 incomplete cases and the first and last case in each labeled family. A fixed rotation selected model and seed, giving 32 outputs per arm, 24 per model, and 48 per seed; this was a coverage sample, not a probability sample or a within-model arm comparison. Supplement S3 records the exact selection rule, unchanged outputs, and deterministic per-output flags. Among the 48 incomplete-case outputs, seven met the registered safe-abstention definition: four returned empty evidence lists and three returned nonempty lists. Twenty-one were registered valid commitments; the other 20 were invalid but contained literal non-null fault_id text. Among 48 labeled-case outputs, 17 met exact diagnosis, including five with partial gold-event coverage. Of 33 length-stopped sample outputs, 28 already repeated complete evidence IDs and three others contained non-visible IDs. Appending tokens alone could not make those 31 prefixes valid and coherent, although a different budget could change generation. These diagnostics concern literal output structure and frozen reference-set overlap, not qualitative relevance. They do not replace registered endpoints or estimate full-journal prevalence.

The numerical relative gate passed despite low absolute diagnosis and abstention rates. This finite suite comprises eight labeled templates and four incomplete templates, with two seeds and four fixed builds. The missingness cue, repeated mechanisms, distractor segregation, and unestablished exhaustive ambiguity construction limit its interpretation. Shared model builds or runtime infrastructure in related evaluations do not constitute additional independent evidence.

For interface evaluation, the result supports measuring output validity alongside classification and abstention when reordering a fixed event set. Stronger reasoning claims would require a new prospective design that varies missingness cues independently of identifiability, retains verifiable latent alternatives, interleaves decoys independently of time order, and includes legitimate controls and novel mechanisms. Longer output budgets would also need a separately specified experiment. These additions are not part of the present evidence.

CONCLUSION

Under the frozen 256-token response budget, trace ordering increased exact diagnosis from 187/640 (29.22%) to 248/640 (38.75%). Safe abstention changed from 21/128 (16.41%) to 23/128 (17.97%), satisfying the registered numerical joint rule. Three builds never safely abstained, and structurally valid commitments rose from 50/128 to 68/128. All 803 length-terminated outputs were invalid; 47 of the net 61 additional correct outputs crossed the validity boundary. The case construction exhibits narrow templates, a perfect missingness cue, systematic decoy segregation, and an unestablished exhaustive two-alternative construction. The defensible finding is end-to-end presentation sensitivity on this recorded synthetic suite. The study does not establish general payment root-cause reasoning, ambiguity recognition, or operational safety.

Funding and competing interests

This research was self-funded. The author declares no competing interests.

REFERENCES

Bendinelli, T., Dox, A., & Holz, C. (2026). *TraceBench: Controlled evaluation of LLM agents for time-series root-cause attribution* [Preprint]. arXiv. https://arxiv.org/abs/2608.27182

Fang, A., Yang, Y., Shang, J., Lu, Q., Xu, J., Wang, R., Zhang, S., Zhang, Y., Yu, B., & He, P. (2026). *OpenRCA 2.0: From outcome labels to causal process supervision* [Preprint]. arXiv. https://arxiv.org/abs/2606.27154

Ghawate, P. (2026). *FinRCA-Bench: Benchmarking evidence retrieval and reasoning for financial AI systems* [Preprint]. arXiv. https://arxiv.org/abs/2608.18534

Gong, A., Choi, K., Agarwal, A., Schechner, J., Huang, R., Agrawal, R., Agarwal, A., & Dwivedi, R. (2026). *ORCA-bench: How ready are language model agents for oncall?* [Preprint]. arXiv. https://arxiv.org/abs/2607.28545v2

Gopal, A., & Krishnan, A. (2026). *How far can root cause analysis go on real-world telemetry data?* [Preprint]. arXiv. https://arxiv.org/abs/2607.13548

Han, Y., Lan, M., & Kilicoglu, H. (2026). When evidence conflicts: Uncertainty and order effects in retrieval-augmented biomedical question answering. In D. Demner-Fushman, S. Ananiadou, K. Roberts, & J. Tsujii (Eds.), *BioNLP 2026* (pp. 630–643). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.bionlp-1.50

Kim, T., Park, W., Yun, H., & Lee, K. (2026). *Why do AI agents systematically fail at cloud root cause analysis?* [Preprint]. arXiv. https://arxiv.org/abs/2602.09937

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. https://doi.org/10.1162/tacl_a_00638

Lu, Q., Fang, A., Xu, J., Shang, J., Zhang, S., Yang, Y., Yan,

X., & He, P. (2026). *Beyond fault localization: A trajectory-level study of LLM agents for microservice root cause analysis* [Preprint]. arXiv. https://arxiv.org/abs/2608.21310

Mohammed, Y., & Lund, B. (2026). Cultural-historical activity theory and AI: Innovating and optimizing financial data retrieval. *American Journal of Financial Technology and Innovation, 4*(1), 75–89. https://doi.org/10.54536/ajfti.v4i1.4187

OpenTelemetry. (2026, January 14). *Traces*. https://opentelemetry.io/docs/concepts/signals/traces/

Presacan, O., Grama, A., Irimină, L., Nik, A., Ojha, J., Thambawita, V., Băcilă, C. I., Ionescu, B., & Riegler, M. A. (2026). *Ask before you diagnose: Safe-Psych, a sequential evaluation benchmark for LLMs in psychiatry* [Preprint]. arXiv. https://arxiv.org/abs/2607.13036

Riddell, E., Riddell, J., Sun, G., Antkiewicz, M., & Czarnecki, K. (2026). Stalled, biased, and confused: Uncovering reasoning failures in LLMs for cloud-based root cause analysis. *In Proceedings of the 2026 IEEE/ACM Third International Conference on AI Foundation Models and Software Engineering* (pp. 172–183). Association for Computing Machinery. https://doi.org/10.1145/3793655.3793732

Serifat, O. A., Igah, R. C., Balogun, K. M., Mensah, G. R., & Odai, E. N. (2025). AI-driven fraud detection in digital banking: ML approach for secure and transparent financial transactions. *American Journal of Financial Technology and Innovation, 3*(1), 177–187. https://doi.org/10.54536/ajfti.v3i1.5168

Xu, R., Li, J., Chen, P., & Xie, Z. (2026). *TELLER: Non-intrusive cross-layer root-cause analysis for LLM inference* [Preprint]. arXiv. https://arxiv.org/abs/2608.01975