



American Journal of Financial Technology and Innovation (AJFTI)

ISSN: 2996-0975 (ONLINE)

VOLUME 4 ISSUE 1 (2026)



PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

Typed Field Support with a Value-Only Deterministic Executor for Synthetic Financial Records: A Multi-Seed Controlled Evaluation

Ayush Ojha^{1*}

Article Information

Received: June 25, 2026**Accepted:** September 09, 2026**Published:** September 30, 2026

Keywords

Financial Reasoning, Language Model, Numerical Reliability, Selective Prediction, Tool-Augmented Reasoning

ABSTRACT

We evaluated whether providing deterministic numerical values to a language model improves supported financial answers without increasing unsupported answer attempts. The internally frozen design crossed 96 synthetic financial bases, four paired evidence conditions, four fixed local model builds, three decoding seeds and three interfaces. All 13,824 scheduled calls completed, yielding 55,296 field decisions. The two principal interfaces shared the same typed support certificate and post-generation gate; only one received a value-only projection from a separately implemented Decimal executor. Relative to TRACE alone, TRACE+Tool increased raw unsupported attempts by 10.1 percentage points (one-sided 97.5% upper bound, 12.2), supported coverage by 16.0 points (lower bound, 14.0), and sound completeness by 86.6 points (lower bound, 85.2). Bounds used stratified resampling of the 96 logical bases. The primary intersection-union test failed because the unsupported-attempt safety margin was exceeded. The broader positive-tool and near-oracle claims also failed. Both TRACE interfaces prevented unsupported emissions by construction, which does not erase the increase in raw attempts. The result identifies a trade-off in this fixed synthetic suite: deterministic value provision improved retained numerical completion while increasing unsupported attempted answers. It does not establish reliability on real financial records or other model populations.

INTRODUCTION

Numerical financial reasoning couples at least three distinct requirements: the source record must contain adequate evidence, the system must select the appropriate operation and units, and the resulting value must be computed correctly. A fluent answer can fail at any of these boundaries. The separation matters because an interface that prevents unsupported answers may still emit wrong values when evidence is adequate, while a calculator can produce exact arithmetic from the wrong or insufficient inputs. Treating either mechanism as a complete reliability solution would therefore hide material failure modes at the support, calculation, or evidence boundary.

Program-aided approaches established that a language model can delegate computation to an interpreter. PAL generates executable programs and delegates their execution to a runtime (Gao *et al.*, 2023), while Program of Thoughts separates language reasoning from external computation and evaluates the approach on mathematical and financial question-answering tasks (W. Chen *et al.*, 2023). Finance-specific neighbors now include evidence-audited program compilation and deterministic execution in DCRC (H. Chen *et al.*, 2026), neurosymbolic orchestration in VADAOchestra (Baldazzi *et al.*, 2026), execution-verified program distillation over financial tables and text (Dong *et al.*, 2026), numerical banking scenarios in BankMathBench (Lee *et al.*, 2026), real-world financial tool retrieval and execution in FinToolBench (Lu *et al.*, 2026), and deterministic tool-grounded multi-agent financial calculation in CIFQA (Parekh *et al.*, 2026).

Selective and evidence-linked generation is also established prior art. PhantomFill studies required-field fabrication under schema pressure (Usman, 2026). Claim-Selective Certification combines mixed evidence, claim-level selection, conflict handling, abstention, and evidence identifiers (Kan, 2026), while ECPO couples candidate ranking to evidence certificates and a deterministic verifier (Hu *et al.*, 2026). CROWN-QA studies when evidence coverage licenses a certified negative rather than an unknown judgment through controlled same-question, same-observed-fact pairs and structured completeness certificates (Min *et al.*, 2026). ExtractBench evaluates field-specific structured extraction, including financial schemas, and distinguishes omission from hallucination (Ferguson *et al.*, 2026). HALT studies partial abstention over factual fragments and reasoning steps (Franzmeyer *et al.*, 2026); calibrated specificity control localizes semantic backoff (Huang *et al.*, 2026); contrastive decoding supports answer-level abstention (Kim *et al.*, 2025); and long-document assistance has combined attribution with abstention (Buchmann *et al.*, 2024).

The experimental comparison holds the typed support certificate, instruction, output schema and deterministic post-generation gate constant between TRACE and TRACE+Tool. TRACE+Tool alone receives a deterministic value-only projection. Each entry contains a value and unit for a field judged supported by a separately implemented checker; it contains no explicit case or condition identifier, support flag, evidence identifier, blocker, primitive, formula or expected label.

¹ 595 Pacific Ave, San Francisco, California, United States* Corresponding author's e-mail: aojha47@gatech.edu

Map membership coincides with supportedness, but that decision is already present in the byte-identical certificate supplied to both TRACE interfaces. The model must still transfer the value and declare the required evidence. This design measures value transfer under a fixed support contract rather than introducing a new calculator, verifier or abstention mechanism.

The prospectively frozen study is a controlled diagnostic over a fixed synthetic suite, not a population study of financial institutions, deployed systems, financial documents, or language models. Its bounded contribution is the joint evaluation of (a) the information-matched TRACE versus TRACE+value-only-projection comparison; (b) paired single-primitive missing, contradictory, and unit-incompatible evidence interventions; and (c) dependency-localized abstention, citation, and unaffected-field value stability across fixed models and seeds. A targeted primary-source refresh through 6 September 2026 did not identify a later record matching this joint design; this remains a dated search boundary rather than a universal priority claim.

Reliability at these boundaries is relevant to auditable automation in consumer-facing financial calculations, internal financial-record processing, and other regulated or high-consequence workflows. That relevance motivates measurement; this synthetic evaluation does not itself establish public benefit, regulatory compliance, production readiness, national-scale impact, or safety for real financial decisions.

The study asks three confirmatory questions:

1. Does adding the deterministic value projection preserve raw unsupported-attempt safety relative to TRACE without the tool?
2. Does it preserve supported coverage?
3. Does it improve sound completeness by more than the prospectively frozen meaningful margin?

All three components, both post-gate structural-zero checks, the zero-failed-call gate, canonical provenance authentication, and every predefined model/task-family/seed robustness gate must pass before any positive tool-benefit claim is permitted. The analyzer reports the three-component primary IUT plus its structural and failure gates separately from the broader robustness-gated positive-tool flag. A distinct, more stringent question compares the tool arm with a direct-executor engineering ceiling. This is not a comparison with an independent language model and cannot by itself justify a deployment claim.

LITERATURE REVIEW

Program-aided reasoning and financial tool execution

PAL delegates model-produced program execution to an external runtime (Gao *et al.*, 2023), and Program of Thoughts applies the separation between language reasoning and computation to numerical tasks that include financial question-answering datasets (W. Chen *et al.*, 2023). The present executor is narrower: it implements

fixed formulas across four task families over a projected visible ledger rather than accepting free-form model-generated programs. That restriction is an experimental control, not a claim that fixed-formula execution supersedes program synthesis.

DCRC combines task-specific evidence acceptance and rejection, program synthesis, deterministic execution, explicit missing and conflicting-evidence bindings, a unit-error rejection code, fixed rounding, and audit trails (H. Chen *et al.*, 2026). It is a close systems comparison; for this comparison, the focal distinctions are the present study's within-TRACE value-only projection and paired localization and unaffected-field stability design. Gold-guided programmatic distillation uses execution-verified programs as supervision for hybrid financial tables and text (Dong *et al.*, 2026). BankMathBench evaluates numerical reasoning and tool augmentation in banking scenarios (Lee *et al.*, 2026), and FinToolBench evaluates financial-tool retrieval and executable use across a large tool inventory (Lu *et al.*, 2026). VADAOchestra combines language-model orchestration with typed tool parameters, dependency-aware prerequisite completion and scheduling, symbolic inference, and replayable logical traces in financial workflows (Baldazzi *et al.*, 2026). CIFQA separates language-model routing, parameter extraction, planning, and response generation from deterministic Python engines for financial query answering (Parekh *et al.*, 2026). CIFQA also provides direct-model comparators with formulas, rate tables, and benchmark instructions; broadly matching available financial information is therefore not new. The distinction here is the within-TRACE control that holds the support contract and gate constant while changing only the value projection.

Several newer financial systems narrow the systems boundary further. AuditFlow combines typed tools, structured evidence, XBRL or taxonomy graphs, required-tool gates, and deterministic numerical and rule verification with evidence trails; it also represents incomplete or ambiguous evidence (Y. Wang *et al.*, 2026). OpenPM applies typed portfolio constraints, a deterministic critic, and point-in-time replay of captured analyst evidence (Cai *et al.*, 2026). AWARE-FX combines exact evidence gates, selective prediction, conservative aggregation, and an audit ledger for corporate foreign-exchange disclosure (Q. Wang, 2026). These works make program-aided financial reasoning, typed controls, deterministic verification, dependency-aware planning, language/computation separation, and auditable orchestration established comparison classes. The present study consequently makes no novelty claim for its executor; it treats the executor as a fixed experimental control for evaluating an independently visible support layer.

MoCA-Agent builds typed catalogs of financial facts, formulas, units and related claims, combines specialist judgments, and synthesizes executable Python programs from the retained evidence (Abdallah *et al.*, 2026). CLAIR-Fin uses a typed financial claim ledger,

deterministic derived quantities, claim-level verification and selective partial answers over multimodal financial documents (Faria *et al.*, 2026). These studies further establish typed financial evidence, executable reasoning and localized abstention as prior art. Their evaluations include system comparisons and component ablations, but neither inspected design isolates the present value-only intervention with a fixed support certificate and gate, paired with controlled defects in individual numerical dependencies.

Selective prediction, certificates, and structured fields

PhantomFill shows that required structured fields can induce fabrication and tests field-level escape behavior (Usman, 2026). ExtractBench provides field-specific evaluation of complex structured extraction, including a broad financial reporting schema, and distinguishes omission from hallucination (Ferguson *et al.*, 2026). These works preclude a claim that field-level schema pressure or structured financial extraction is new.

Claim-Selective Certification already combines mixed evidence, partial or full claim selection, conflict handling, abstention, and evidence-identifier certificates (Kan, 2026). ECPO couples evidence-certified decisions to a deterministic verification cycle (Hu *et al.*, 2026). HALT trains partial abstention over factual fragments and reasoning steps (Franzmeyer *et al.*, 2026), while calibrated claim-level specificity controls how far an answer backs off under uncertainty (Huang *et al.*, 2026). Kim *et al.* (2025) study training-free answer-level abstention, and Buchmann *et al.* (2024) combine source attribution with abstention for long documents. Accordingly, partial abstention, evidence-linked certification, evidence identifiers, deterministic certificate verification, and attribution with abstention are all prior art.

CROWN-QA is especially relevant because it varies evidence coverage while holding the question and observed facts fixed, then asks whether absence supports a certified negative or requires an unknown answer (Min *et al.*, 2026). Its elicited certificate is generated by the model and mapped to the two-label decision, not supplied as deterministic authorization for multiple numerical outputs. TRACE claims neither coverage-sensitive abstention nor structured completeness certificates as new. Its different unit of analysis is a set of simultaneously requested numerical fields with registered typed dependencies; the benchmark also introduces contradictory and unit-incompatible evidence and measures citation behavior and unaffected-field value stability.

Financial evaluation work also separates evidence quality from downstream reasoning. FinRCA-Bench fixes the reasoning model, prompt, and generation settings while varying retrieval and evaluates evaluator-private record-level evidence contracts, typed provenance, and citation-identifier validity (Ghawate, 2026). Its contracts are not the model-visible support authorization used here, and its reported citation metric does not adjudicate semantic

support. The EvidenceLens prototype visualizes coverage, contradiction, and support gaps through a financial claim-evidence matrix (Gu *et al.*, 2026). FinRiskAtlas evaluates static financial-review tasks and offline replay of recorded evidence-acquisition decisions (Zhong *et al.*, 2026). Replayable Financial Agents measures trajectory determinism, decision determinism, and evidence-conditioned faithfulness across repeated financial-agent executions (Khatchadourian, 2026). Its fault scenarios include data-quality stress, but the pilot stress columns are explicitly projections rather than measured core results. These studies establish evidence-state controls, provenance-oriented evaluation, and replay as adjacent comparison classes; the present benchmark instead registers per-field calculation dependencies and paired primitive defects within each logical base.

Mohammed and Lund (2026) evaluate a financial-document chatbot using context recall, faithfulness, context utilization, and context precision. Ogunsusi *et al.* (2026) review generative AI and advanced analytics for financial modeling, valuation, and strategic decision-making, including human oversight and reliability risks. The present experiment narrows its evaluation to support decisions and numerical completion under fixed visible synthetic records; it does not evaluate document retrieval or organizational decision outcomes.

Information-matched comparison and bounded gap

The present benchmark manipulates evidence sufficiency within matched synthetic records. Each logical base has complete, single-primitive missing, contradictory, and unit-incompatible variants. TRACE and TRACE+tool receive the same support certificate, instruction, output schema, and gate; only the deterministic value-only projection differs. Hidden expected labels remain outside both the model-visible contract and the independent executor.

A targeted primary-source refresh through 6 September 2026 included MoCA-Agent and CLAIR-Fin alongside DCRC, AuditFlow, CROWN-QA, CIFQA, FinRCA-Bench and Replayable Financial Agents. These sources cover deterministic financial execution, typed evidence, certificates, component ablations, fixed-reasoner retrieval comparisons and replay-based evaluation. Among these inspected evaluations, none matched the complete joint design studied here: visible per-field typed-support authorization, an information-matched value-only execution control, paired missing, contradictory and unit-incompatible primitive interventions, and dependency-localized abstention, citation and unaffected-field stability measurements. This is a bounded comparison of identified sources, not a universal priority claim.

Experimental synthesis and contribution boundary

The comparison is organized around three separable functions that are often combined in an end-to-end financial question-answering system. The support layer decides whether the visible record contains an admissible

dependency chain for each requested field. The execution layer applies a fixed arithmetic mapping to visible inputs. The post-generation gate checks the model's declared status, blockers, evidence identifiers, and citations before a value can be emitted. Separating these functions permits a more specific controlled within-suite treatment contrast than whether a general tool-using agent outperforms a non-tool agent. The intervention changes the availability of a value-and-unit projection while holding the support certificate, response schema, decoding controls, and gate constant between the two TRACE arms.

This separation also constrains interpretation. A gain in numerical accuracy would not show that the model discovered better evidence, because the certificate is unchanged. A loss of coverage would not establish that tools are inherently unsafe, because the executor and projection are deliberately narrow and the result would be conditional on the tested interface. A structural zero in post-gate unsupported emission would demonstrate the operation of a deterministic filter, not the absence of unsupported attempts by the model. The registered endpoints therefore distinguish attempted behavior, accepted behavior, coverage, and correctness rather than collapsing them into a single answer-accuracy score.

Field-level typing matters because a single record can support some requested values while leaving others missing, contradictory, or unit-incompatible. Answer-level abstention would treat that mixed state as one decision and could hide collateral refusal or unsupported partial completion. The benchmark instead retains four simultaneous field decisions, dependency-localized interventions, citation checks, and unaffected-field diagnostics. This design does not make typed support a truth certificate: the certificate can only describe the registered visible dependencies, and a supported field may still receive an incorrect value. The direct executor, model declaration, and gate therefore remain independently observable components.

The resulting contribution is a controlled measurement design and evidence package rather than a new tool primitive. It can yield an informative result even if the primary IUT fails: the fixed denominators and localization diagnostics can identify whether the tested projection changes unsupported attempts, coverage, transfer fidelity, or value correctness. Conversely, a positive fixed-suite result would remain evidence about this checked-in generator, interface contract, and model set, not a license to infer behavior for production systems or a population of financial tasks.

MATERIALS & METHODS

Prospective status and protocol lineage

The v1.7.2 operational-recovery study was frozen before any v1.7.2 pilot or main generation request. It is prospectively pre-outcome with respect to its own scored rows, but it followed disclosed observation of the v1.7 and v1.7.1 pilot records. The canonical sources for this manuscript's methods are `configs/study-v1.7.2.json`,

`docs/11-v1.7-prospective-tool-protocol.md`, `docs/12-v1.7.1-prospective-schema-bounds-protocol.md`, and `docs/13-v1.7.2-operational-recovery-protocol.md`. The required annotated tag is `trace-finance-v1.7.2-freeze`.

The excluded v1.7 operational pilot contained 575 valid calls and one authenticated call terminated at the fixed 768-token ceiling. The failed field response repeated invented evidence identifiers. The machine gate failed, and no v1.7 main run exists. Version 1.7.1 was defined before its replacement-pilot request and changed only provider-schema bounds for the two previously unbounded response arrays. That replacement pilot completed and passed its registered machine gate, but the registered audit private key could not be unlocked; no v1.7.1 main was executed.

Version 1.7.2 is therefore a post-pilot-informed operational recovery, not an independent preregistration. Before its one-shot pilot, it rotated the pilot and main seeds, refreshed the observed NVIDIA driver version from 595.97 to 610.74, registered a new dedicated audit key, and added recovery and local compare-and-swap controls. It did not change the benchmark, hypotheses, endpoints, margins, arms, model digests, prompts, decoding, cohort, or analysis rules, and no v1.7 or v1.7.1 pilot response enters the v1.7.2 main analysis.

The protocol and analysis plan were frozen before the first scored v1.7.2 pilot request. This outcome-neutral manuscript scaffold and its extraction mapping were sealed later: after the excluded v1.7.2 pilot passed its registered authorization gate, but before any v1.7.2 main outcome existed. That timing does not make the manuscript an independently registered preregistration. No pilot row is eligible for a main estimate or decision. The benchmark and typed support layer had been evaluated in the earlier v1.6 study; the present freeze is prospective for the new multi-seed tool-comparison responses, rather than a first evaluation of previously unseen cases.

For each requested field in all arms, the schema restricts `evidence_ids.items.enum` to the sorted distinct evidence identifiers in the model-visible case ledger and sets `maxItems` to that vocabulary size. `blocking_keys.items.enum` is restricted to the sorted primitive-key universe for the visible task family with the corresponding `maxItems`. Empty arrays remain allowed. These are whole-case public vocabularies, not per-field dependency closures, support decisions, expected citation counts, or hidden labels. Evidence rows are projected to exactly `evidence_id`, key, unit, and value before prompt, schema, certificate, or tool construction.

Benchmark and paired evidence conditions

The benchmark contains 96 logical `base_id` clusters, balanced across four synthetic financial task families with 24 bases per family. Each base has four paired evidence conditions:

- complete;
- missing;
- contradiction; and

- unit_mismatch.

Each case schedules four requested fields. The checked-in benchmark contains 384 condition-specific records. The fixed design therefore supports within-base comparisons across evidence interventions, but the 384 records are not 384 independent experimental units. The resampling unit remains the 96 logical bases.

The deterministic generator used seed 20260731 and allocated 24 bases to each of four families. Installment-loan primitives were principal (\$2,000–\$50,000 in \$100 steps), annual rate (2.50%–19.99%), term (12, 24, 36, 48, or 60 months), and origination fee (\$0–\$750). Savings-plan primitives were initial deposit (\$500–\$20,000 in \$100 steps), monthly contribution (\$25–\$1,000 in \$5 steps), annual rate (0.50%–8.50%), and duration (12, 24, 36, 60, 84, or 120 months). Card-statement primitives were opening balance (\$0–\$8,000), purchases (\$0–\$4,000), payments (\$0–\$3,500), credits (\$0–\$500), fees (\$0–\$150), and interest (\$0–\$250). Merchant-settlement primitives were gross sales (\$1,000–\$50,000), refunds (\$0–\$2,500), chargebacks (\$0–\$1,000), processing rate (1.50%–3.99%), per-transaction fee (\$0.10–\$0.75), and transaction count (10–2,000).

Each family returned four registered derived fields. The installment family used the standard fixed-rate amortizing-payment equation, followed by total payments, finance charge, and fee-inclusive cost. The savings family used a monthly compound factor and ordinary-annuity contribution factor, followed by contributed principal, ending balance, and interest earned. The card family used additive payments-and-credits, charges, net activity, and ending-balance roll-forward equations. The merchant family used variable and fixed processing fees, dispute deductions, and net settlement. Currency outputs and currency intermediates were rounded half up to cents; the reported compound factor was rounded half up to six decimal places. The supplementary materials provide the study protocol and synthetic benchmark; a separate source archive contains the dependency and calculation implementations.

The perturbed primitive was fixed by family: origination fee, savings annual rate, card interest, or merchant chargebacks. The missing condition removed that row.

The contradiction condition retained the original and appended a second row whose value increased by the larger of 0.01 and 10% of the original absolute value. The unit-mismatch condition retained the value but replaced its unit with a fixed incompatible-unit token. Other primitives were unchanged. The 96-base cohort was retained after the prior v1.6 study produced approximate 1.5–2.1 percentage-point interval half-widths; this is disclosed precision context for the synthetic design, not a prospective power calculation or a claim about a financial population.

Model and decoding controls

The frozen model order was qwen3:8b (8.2B parameters, Apache License 2.0), gemma3:4b (4.3B parameters, Gemma Terms of Use), llama3.1:8b (8.0B parameters, Llama 3.1 Community License), and phi4-mini:3.8b (3.8B parameters, MIT License). All four evaluated builds used GGUF Q4_K_M quantization. Supplement S1 and the recorded configuration supply the complete provider-manifest SHA-256 digests; these pin the evaluated artifacts even if a public Ollama tag later changes. The frozen runtime records Ollama 0.32.14, Python 3.11.14 and fastjsonschema 2.22.1 on an NVIDIA GeForce RTX 5090 Laptop GPU. Model weights were accessed under their respective licenses and are not redistributed with the manuscript materials.

The main study uses three prospectively frozen seeds: 20260830, 20260831, and 20260901. Temperature is 0.2, the prediction ceiling is 768 tokens, and thinking mode is disabled. Interface order follows a seed-rotated three-period Latin square so each interface occupies each within-case ordinal position. The complete main schedule contains 13,824 language-model calls and 55,296 scheduled field decisions. These are design counts, not observed completion counts.

An excluded v1.7.2 pilot crossed four bases (one per family), all evidence conditions, four model digests, three arms, and three disjoint pilot-only seeds (20260827–20260829), yielding 576 scheduled calls. Its responses are excluded from the main analysis; latency informed operations only and could not alter endpoints, margins, seeds, decoding, or the frozen design. The final execution-

Table 1. Frozen study design (not outcomes)

Design element	Prospectively frozen value
Logical base clusters	96
Task-family strata	4
Bases per stratum	24
Conditions per base	4
Fixed model digests	4
Main decoding seeds	3
LLM interfaces	3
Scheduled main calls	13,824
Scheduled main field decisions	55,296
Bootstrap replicates	10,000

integrity report authenticates its machine gate and dedicated artifact audit without treating pilot outcomes as tool-benefit evidence. Figure 1 summarizes the study flow; Table 1 lists the frozen design.

Interfaces

The field arm is the model-only field-selective interface. The trace arm adds a typed support certificate, a common permissive TRACE response schema, and a deterministic post-generation gate. The trace_tool arm uses the byte-identical TRACE support information, instructions, response schema, and post-generation gate, plus a value-only deterministic projection for supported fields.

The tool projection has the exact conceptual form field $\rightarrow \{value, unit\}$. It contains no explicit case or condition identifier, support or status flag, evidence, blocker, primitive, formula, calculation metadata, or tool hash. Because entries exist only for supported fields, map membership redundantly encodes the support decision

already disclosed in the common certificate; it adds no new support information beyond that certificate. The model must transfer the value and separately declare citations from the same support certificate used by the no-tool TRACE arm. The gate does not replace a wrong numeric value, leaving transfer failures observable.

The direct executor reads only the visible ledger, public case identity, requested fields, and output units. It does not import the benchmark calculator or read hidden expected labels. The formulas across the four task families are separately implemented with Python Decimal. The executor rejects non-finite values; fractional, zero, or negative counts; zero-denominator and non-finite formula domains; unknown fields or units; and disagreement with the independent TRACE support checker.

An oracle direct-executor arm is constructed only in analysis as an engineering ceiling. It does not invoke a language model and is not counted as another model or seed observation.

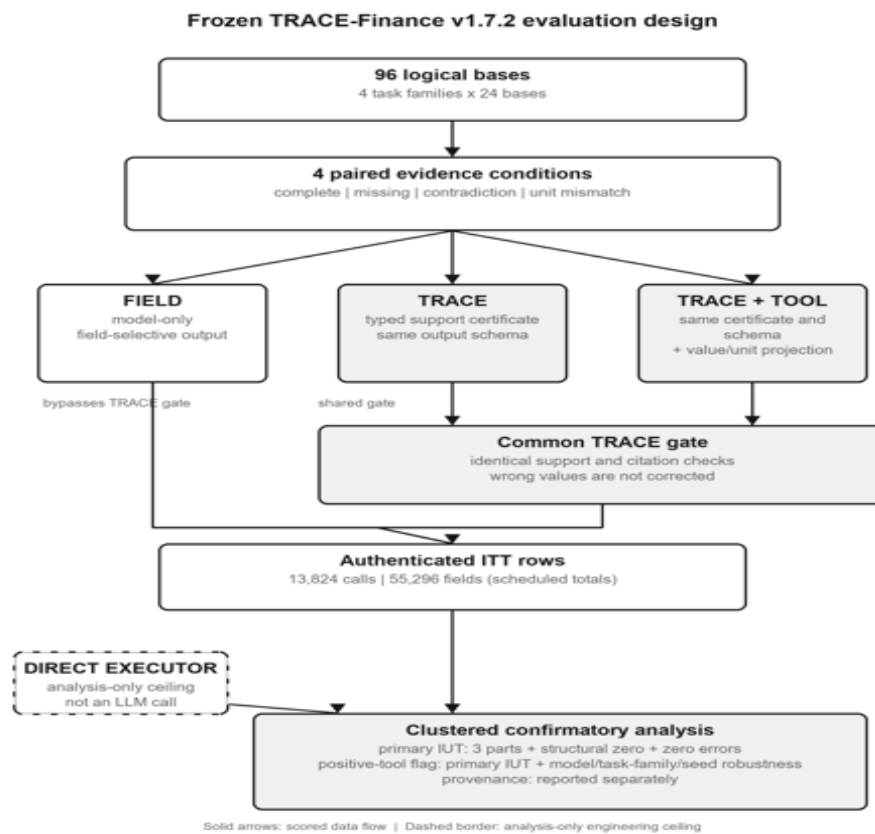


Figure 1: Frozen TRACE-Finance v1.7.2 evaluation design. Each of 96 logical synthetic financial bases produces four paired evidence conditions. The same case is evaluated by the model-only field interface, the typed TRACE interface, and the TRACE interface with a value-only deterministic Decimal projection. All TRACE outputs pass through the same support/citation gate. Analyses retain all repeated observations within the logical-base cluster. The primary intersection-union decision combines three statistical components with structural-zero and zero-error gates; the positive-tool flag additionally requires the model/task-family/seed robustness screen, and canonical provenance authentication is reported separately from the statistical components. The direct executor used as an engineering ceiling is not a language-model arm.

Endpoints

The prospectively frozen endpoints are:

- raw unsupported-attempt rate;
- post-gate unsupported emission rate (UER);

- supported coverage (SC);
- sound completeness (SComp), defined as correct supported values divided by all scheduled supported fields;

- value accuracy when emitted (VAE), a secondary conditional diagnostic;
- exact tool-transfer accuracy;
- citation completeness and citation purity;
- condition and intervention localization;
- collateral abstention and unaffected-field value changes;
- unique-call latency and token summaries; and
- exact three-seed agreement over the full normalized declaration.

For a complete main run, each language-model arm contributes 18,432 scheduled field decisions: 10,656 supported and 7,776 unsupported. Raw and post-gate UER use all unsupported fields; SC and SComp use all supported fields; VAE uses only supported fields that the corresponding arm emits and is therefore an arm-specific, selection-sensitive conditional proportion. A value is scored correct when its absolute error is at most \$0.01 for USD or 0.000001 for other registered units. Tool-transfer accuracy uses every supported trace_tool field. Citation completeness divides required evidence identifiers cited by all required identifiers among supported answered fields. Citation purity divides admissible citations by all citations among those fields; because the response schema restricts every citation to the whole-case visible evidence-ID vocabulary, defined purity is a schema-conformance invariant of 1.0 rather than evidence that the model selected the field's required citations.

Collateral abstention is the fraction of raw-answered supported TRACE fields removed by the deterministic gate. For each non-complete condition, unaffected-field diagnostics pair only fields supported in both the complete and perturbed record. Value-change rates condition on both members being answered, while complete-answer-loss rates condition on the complete member being answered. Every conditional denominator is reported explicitly; a zero denominator remains non-estimable.

A raw answer attempt is defined before gating as a declared answer status or a nonempty value. Citations without either condition do not constitute an answer attempt. Applying this effective-answer definition to all arms prevents an inconsistent status/value pair from evading the raw safety endpoint. Post-gate UER of zero in the TRACE arms is a structural implementation invariant, not an empirical efficacy finding.

Intention-to-treat accounting

SComp uses a fixed intention-to-treat denominator. A malformed, missing, schema-invalid, telemetry-invalid, provider-invalid, or otherwise failed scheduled call creates one authenticated failure row. All four fields in that call score incorrect. Errors are counted by the unique (case_id, model, seed, method) schedule key, not by the number of fields. The runner does not retry an ordinary per-call failure and continues the Cartesian schedule. Pilot approval requires zero errors, and any positive tool-benefit or near-oracle claim also requires zero failed calls.

Confirmatory estimands

The confirmatory raw-UER, SC, and SComp effects are paired pooled risk differences over the complete fixed benchmark/model/seed suite. VAE contrasts instead subtract arm-specific conditional proportions whose emitted-field risk sets can differ, so they are selection-sensitive secondary diagnostics, not common-denominator paired effects. The primary intersection-union test requires all of the following. The intersection-union framing follows the equivalence-testing treatment of Berger and Hsu (1996):

1. the one-sided 97.5% upper bound for rawUER(trace_tool) - rawUER(trace) is below +0.005;
2. the one-sided 97.5% lower bound for SC(trace_tool) - SC(trace) is above -0.02; and
3. the one-sided 97.5% lower bound for SComp(trace_tool) - SComp(trace) is above +0.10.

Each component is evaluated at one-sided alpha 0.025. The primary IUT flag is also gated on zero post-gate unsupported emissions in both TRACE arms and zero failed calls. The broader positive tool-benefit claim is a separate machine flag that additionally requires the prospective model/task-family/seed robustness screen. The manuscript reports both flags and does not fold robustness into the primary IUT.

A separate near-oracle claim requires the lower bound for SComp(trace_tool) - SComp(oracle) to exceed -0.02. It additionally requires its predefined Holm decision, zero post-gate tool-arm UER, both predefined raw safety and coverage bounds, and zero failures. VAE for TRACE versus the field arm is retained as a secondary sensitivity comparison against a -0.05 noninferiority margin; it is not an exact replication of an earlier temperature-zero study. The prospectively frozen three-member secondary family uses Holm correction at familywise alpha 0.025. A conditional VAE contrast with no eligible emitted values is non-estimable, remains in the family with $p = 1$, and cannot support a claim. The family is never reduced after outcome inspection. The sequentially rejective adjustment follows Holm (1979).

Uncertainty and robustness

Exact point contrasts describe the complete fixed suite. The analysis uses 10,000 bootstrap replicates. Each replicate nonparametrically resamples 24 of the observed logical base clusters with replacement within each of the four generator-defined task-family strata, retaining every associated condition, model, seed, field, and interface. It does not generate new cases. The bootstrap seed is 20260820.

The resampling implementation follows the nonparametric bootstrap principle (Efron, 1979) with clustered, stratified resampling appropriate to the nested design; Davison and Hinkley (1997) provide the broader bootstrap framework.

Bootstrap bounds describe sensitivity to empirical resampling of observed base clusters within the four generator-defined strata. They are not population intervals

for real financial documents, institutions, production decisions, arbitrary tools, or a model population. The predefined secondary p-values use a null-centered, task-stratified base-cluster bootstrap; they are not exact randomization p-values.

The robustness screen requires SComp improvement greater than 10 percentage points in at least three of four fixed model digests, at least three of four designed task families, and every prospectively frozen main seed. These are fixed-design screens rather than population inference.

Execution integrity and provenance

The canonical runner checks the exact repository root, tagged source bytes, tracked-file state, benchmark hash, model order and digests, runtime and hardware identity, full-GPU residency, tool-bundle recomputation, and complete schedule before provider calls. An exclusive operating-system lock prevents concurrent pilot or main invocations. This lock serializes canonical study invocations; it does not reserve the GPU against unrelated applications.

The journal durably records started before each request and completed after the hashed row. A started request without a durable row after an interruption is not resent; it becomes an authenticated interrupted_ambiguous intention-to-treat failure. A durable response missing only its completion event can receive that event without a provider resend.

Run locks bind source, configuration, implementation, benchmark, tool bundle, request protocol, schedule, models, runtime, hardware, decoding, and pilot approval. Response rows bind the provider envelope, raw response, telemetry, pre-gate declaration, post-gate fields, interventions, prompt/schema contract, and tool hashes where applicable. Ordinary authenticated response failures remain observations. Source drift, runtime drift, tamper, impossible journal state, noncanonical substitution, or an incomplete schedule closes the run as invalid.

Before the main run, the excluded pilot passed its registered machine gate, and the frozen artifact hashes and detached signature were validated.

After the run, a separate automated recomputation authenticated the canonical bundle and reproduced the registered metrics from fresh artifact snapshots.

Data, software, and study scope

The benchmark is synthetic and contains no recruited participants, identifiable private or customer records, or production financial data. The study used locally executed inference over synthetic inputs.

Validity controls and reporting discipline

Construct validity is addressed by reporting several noninterchangeable quantities. Raw unsupported attempts capture model behavior before the gate; post-gate UER captures what the interface permits to leave the

gate; supported coverage records retained answerability; SComp uses the full supported-field denominator; and VAE is explicitly conditional on emission. Citation completeness, tool-transfer accuracy, and localization diagnostics provide failure descriptions but do not replace the registered confirmatory estimands. Citation purity is retained as a schema-conformance check and is mechanically 1.0 whenever defined because out-of-vocabulary identifiers fail validation. This separation prevents a mechanically enforced post-gate zero or purity one from being presented as evidence that the model itself ceased making unsafe attempts or always chose required evidence.

Internal-validity controls include paired evidence interventions within each logical base, identical support information and response schemas for the two TRACE arms, a shared post-generation gate, seed-rotated interface order, and an executor that cannot read hidden expected labels. The value-only projection is intentionally narrow: it contains no explicit case identifier or support flag, and its field membership adds no support information beyond the common certificate, although values can still be distinctive of a record. These controls isolate the tested interface difference more tightly than an unmatched comparison, although they cannot eliminate all order, prompt, or provider-runtime interactions.

Statistical-conclusion controls include intention-to-treat accounting for every scheduled field, task-stratified logical-base resampling, a prospectively fixed three-part intersection-union decision, a complete three-member Holm family, and predefined model, task-family, and seed screens. Conditional estimates remain non-estimable (serialized as JSON null) when no eligible denominator exists; a registered non-estimable secondary member remains in the Holm family with $p = 1$. The frozen reporting rule retained every registered component and diagnostic regardless of direction, without omitting a failed component, reducing the Holm family, reinterpreting a two-sided contrast as improvement, or substituting a design total for an authenticated observed count.

External validity remains deliberately limited. The four task families are synthetic, structurally balanced, and generated under a fixed specification; the four model digests, three seeds, local runtime, formulas, and evidence interventions are not random samples from real institutions or deployed systems. Bootstrap bounds quantify sensitivity to resampling from the fixed generator strata, not uncertainty over an institution, document population, model population, or future provider version. Accordingly, the fixed-suite interpretation includes the operational, rights, and provenance limitations described here.

RESULTS AND DISCUSSIONS

All values below come from authenticated canonical artifacts and were reproduced by a separate recomputation from fresh snapshots. That check did not repeat model inference.

Execution integrity

The authenticated schedule contained 13,824 calls; all completed and none failed. Each call contributed four scheduled field decisions. Table 2 summarizes

the execution checks; Supplement S1 retains the full source tags, hashes, metric paths and observed runtime values, and the recorded evidence files permit numerical reconstruction.

Table 2: Execution and provenance checks

Check	Observed value	Interpretation
Complete schedule	13,824/13,824 calls	No failed calls
Field accounting	55,296 decisions	All scheduled fields retained
Source and runtime	Frozen v1.7.2 contract authenticated	Full identifiers in S1
Recorded metrics	Separate snapshot recomputation	No repeated inference
Post-gate unsupported emission	TRACE 0/7,776; TRACE+Tool 0/7,776	Structural filter invariant

Arm-level endpoints

Table 3 reports arm-level endpoints with their eligible numerators and denominators. Percentages are rounded to one decimal place; all decision calculations use the unrounded recorded values.

All arm-level fractions are descriptive; no separate arm-

level interval is claimed. Value accuracy when emitted is conditional on each arm's emitted values. The zero unsupported-emission rates in the TRACE arms are enforced by the gate. Exact tool transfer requires equality of the recorded decimal strings, whereas sound completeness uses the registered numeric tolerance.

Table 3. Arm-level outcomes as numerator/denominator (percentage)

Endpoint	Field	TRACE	TRACE+Tool
Raw unsupported attempts	3,529/7,776 (45.4%)	3,799/7,776 (48.9%)	4,588/7,776 (59.0%)
Post-gate unsupported emissions	3,474/7,776 (44.7%)	0/7,776 (0.0%)	0/7,776 (0.0%)
Supported coverage	5,712/10,656 (53.6%)	8,441/10,656 (79.2%)	10,148/10,656 (95.2%)
Sound completeness	589/10,656 (5.5%)	871/10,656 (8.2%)	10,101/10,656 (94.8%)
Value accuracy when emitted	589/5,712 (10.3%)	871/8,441 (10.3%)	10,101/10,148 (99.5%)
Exact value transfer, raw	Not applicable	Not applicable	10,191/10,656 (95.6%)
Exact value transfer, post-gate	Not applicable	Not applicable	10,099/10,656 (94.8%)

Primary intersection-union decision

The TRACE+Tool-minus-TRACE raw unsupported-attempt-rate difference was +10.1 pp with a one-sided upper bound of +12.2 pp; supported coverage was +16.0 pp with a lower bound of +14.0 pp; and sound completeness was +86.6 pp with a lower bound of +85.2 pp. The primary IUT did not pass (Table 4). The separate robustness-gated positive tool-claim flag did not pass.

A post-hoc check of the transfer discrepancy found two numerically equivalent string differences: 3420.0 versus 3420.00 and 13240.0 versus 13240.00. These explain the difference between 10,099 exact transfers and 10,101 correct emitted values. Of the 10,656 supported tool-arm fields, 47 had incorrect emitted values and 508 were not emitted. This descriptive accounting clarifies the endpoints without changing any registered decision.

Table 4: Primary intersection-union test

Component	Required bound	Difference	97.5% bound	Decision
Raw unsupported attempts	Upper < +0.5 pp	+10.1 pp	Upper +12.2 pp	Fail
Supported coverage	Lower > -2.0 pp	+16.0 pp	Lower +14.0 pp	Pass
Sound completeness	Lower > +10.0 pp	+86.6 pp	Lower +85.2 pp	Pass

Secondary and robustness analyses

Table 5 reports the complete prospectively frozen three-member Holm family. Table 6 summarizes fixed-design robustness and Table 7 summarizes selected aggregate diagnostics; all predefined diagnostic cells, conditional denominators and exact metric paths remain in Supplement S1. The third secondary member is the

prespecified two-sided TRACE+Tool-minus-TRACE VAE comparison against zero. Its positive signed estimate and rejection indicate a conditional difference on arm-specific emitted-field denominators. They do not establish a common-denominator accuracy improvement or rescue the failed primary test.

Table 5: Near-oracle and Holm-adjusted secondary family

Comparison	Criterion	Estimate / bound	Raw p / adjusted p	Interpretation
Near-oracle SComp	Tool – oracle; lower > –2.0 pp	–5.2 pp; lower –6.1 pp	1.0000 / 1.0000	Failed; additional claim gates also apply
TRACE – Field VAE	Lower > –5.0 pp	+0.0 pp; lower –1.1 pp	0.0001 / 0.0003	Conditional noninferiority
Tool – TRACE VAE	Two-sided difference from zero	+89.2 pp; 95% interval [87.9, 90.5] pp	0.0001 / 0.0003	Conditional difference; observed direction positive

VAE comparisons use different emitted-field denominators. Holm correction retains all three comparisons at familywise $\alpha = 0.025$; full ranks and stepwise thresholds appear in the recorded metrics. The two p-values shown as 0.0001 are rounded finite-bootstrap values, not zero probabilities. The VAE difference interval is two-sided 95%; the noninferiority bounds are one-sided 97.5%.

All four model builds, all four task families and all three seeds exceeded the fixed 10-point SComp threshold. The required counts were at least 3/4 models, at least 3/4 task families and 3/3 seeds. These screens concern completeness only; the separate raw-safety failure still prevents the positive-tool claim. Condition-specific contrasts and full model digests are retained in S1.

Table 6: Sound-completeness differences across the fixed design

Grouping	Slice	SComp difference	Criterion
Model build	Qwen3 8B	+79.3 pp	>10 pp; passed
Model build	Gemma3 4B	+94.5 pp	>10 pp; passed
Model build	Llama3.1 8B	+89.7 pp	>10 pp; passed
Model build	Phi4-mini 3.8B	+83.0 pp	>10 pp; passed
Task family	card statement	+67.7 pp	>10 pp; passed
Task family	installment loan	+95.4 pp	>10 pp; passed
Task family	merchant settlement	+79.5 pp	>10 pp; passed
Task family	savings plan	+99.4 pp	>10 pp; passed
Main seed	20260830	+86.9 pp	>10 pp; passed
Main seed	20260831	+86.3 pp	>10 pp; passed
Main seed	20260901	+86.7 pp	>10 pp; passed

Citation completeness counts required identifiers among supported answered fields. Citation purity was 100% whenever defined, a consequence of the schema's whole-case evidence vocabulary rather than semantic support. Agreement uses the full normalized field declaration, not correctness. Latency describes 4,608 observed calls

per arm on a shared host; the 5,024-second maximum is retained and the mean is not an isolated-device speed estimate. Supplement S1 includes every condition-level intervention, unaffected-field comparison, token summary and family-level transfer result.

Table 7: Aggregate descriptive diagnostics

Diagnostic	Field	TRACE	TRACE+Tool
Raw citation completeness	15,880/16,139 (98.4%)	24,751/24,801 (99.8%)	29,341/29,799 (98.5%)
Post-gate citation completeness	15,880/16,139 (98.4%)	24,741/24,741 (100%)	29,227/29,227 (100%)
Supported answers removed by gate	Not applicable	10/8,451 (0.1%)	102/10,250 (1.0%)
Exact three-seed agreement, raw	1,237/6,144 (20.1%)	1,998/6,144 (32.5%)	3,757/6,144 (61.1%)
Exact three-seed agreement, post-gate	1,237/6,144 (20.1%)	2,764/6,144 (45.0%)	4,729/6,144 (77.0%)
Call latency, median / 90th percentile (s)	3.553 / 4.558	3.519 / 4.310	3.542 / 4.305
Call latency, mean / maximum (s)	3.673 / 15.503	3.603 / 18.831	4.701 / 5,024.260

Interpretation

The deterministic value projection substantially increased retained numerical completion: TRACE+Tool returned correct supported values for 10,101 of 10,656 eligible fields, compared with 871 for TRACE. However, its raw unsupported attempts increased from 3,799 to 4,588 of 7,776 unsupported fields. The additional 789 attempted answers correspond to a 10.1-point increase, with an upper bound of 12.2 points against the prespecified +0.5-point margin, so the joint primary criterion failed despite the completeness and coverage gains. Reporting only the gated outputs would conceal this change in attempted behavior.

The deterministic gate blocked unsupported emissions in both TRACE interfaces. This property depends on the benchmark support contract and should be evaluated separately from the language model's willingness to answer. Providing correct values for supported fields did not make the model abstain reliably on unsupported fields. The experiment measures this joint response to the interface; it does not isolate a cognitive explanation for the extra attempts. The separately gated near-oracle claim failed, and the conditional VAE results do not establish the broader positive-tool claim.

The direct executor is an informative engineering baseline: when admissible inputs and formulas are already known, a deterministic calculation does not require a language model to reproduce its output. Here the model adds a transfer and declaration step, which remained imperfect. The study therefore supports separating evidence sufficiency, calculation and emitted-answer checks when evaluating such interfaces. Whether a language model adds value in a larger workflow that requires retrieval, entity resolution or operation selection remains outside this experiment.

Limitations

The benchmark is synthetic and structurally balanced. It does not reproduce the distribution, ambiguity, formatting, temporal change, confidentiality, adversarial behavior, or legal constraints of live financial records. The inference target is the checked-in generator and fixed strata, not a population of financial cases.

The model suite contains four fixed local digests. The design cannot estimate performance over a population of language models, commercial APIs, future weights, quantization choices, or deployment configurations. Low-temperature multi-seed runs measure sensitivity within the frozen decoding contract, not all stochastic behavior. The execution host was not exclusively reserved. Outcome-blind operational observations during the main run identified a concurrent interactive GPU workload and a period of high system committed-memory utilization; later observations showed lower memory allocation. These observations do not establish causal effects on model behavior, timing, or call failures. Any latency or resource summaries characterize the observed shared-host execution rather than isolated-device performance.

The registered intention-to-treat and failure-handling rules remain unchanged; operational interference is not grounds for selectively removing or repeating calls.

The deterministic executor covers fixed formulas across four task families and receives contract-defined visible inputs. It is not a general financial calculator, retriever, planner, verifier, or arbitrary program-execution environment. Its engineering-ceiling role is deliberately narrow.

Support checks and schema bounds depend on benchmark-defined typed primitives and public vocabularies. Real documents may contain unresolved entity, currency, period, provenance, and unit ambiguities outside this representation. The structural post-gate UER invariant tests implementation conformance under the contract, not empirical efficacy or semantic safety beyond the contract.

The bootstrap assumes exchangeability of logical bases within the four synthetic task-family strata. The predefined p-values are null-centered bootstrap diagnostics, not exact randomization tests. Multiple subgroup screens are claim gates and descriptive diagnostics, not population hypothesis tests.

The v1.7 pilot failure is evidence that response-length and schema behavior can invalidate a run even when most scheduled calls complete. Version 1.7.1 repairs bounded response arrays without changing the scientific design, but the repair does not prove that other provider or schema failures are absent.

The one-shot attempt ledger and compare-and-swap guard are local controls with no Git remote anchor or external append-only timestamp. They deter accidental replacement runs and bind the disclosed local chronology, but they cannot prevent an operator from deliberately rewriting local Git objects or refs. Neither the local timestamp nor the detached audit signature establishes an external trusted time source or human identity.

CONCLUSION

Within the fixed suite, the value-only projection raised supported coverage by 16.0 pp and sound completeness by 86.6 pp, but it also raised the raw unsupported-attempt rate by 10.1 pp. Because the registered raw-safety component failed, the primary intersection-union test and broader positive tool-benefit claim were not established. The separately gated near-oracle claim also failed. The TRACE-versus-field conditional value-accuracy-when-emitted noninferiority test passed. The TRACE+Tool-versus-TRACE conditional contrast was +89.2 pp on arm-specific emitted-field denominators; its two-sided Holm result establishes a difference with positive observed direction, not a standalone improvement claim. Neither secondary result rescues the primary failure. Both TRACE arms had zero post-gate unsupported emissions by construction; that structural invariant does not erase the raw-attempt harm. The bounded lesson is that this tested deterministic-value interface increased retained correct completion while simultaneously eliciting more

unsupported raw attempts. No deployment safety, regulatory compliance, real-record performance, or generalization claim is made.

Data Availability

Supplement S1 contains the complete detailed result tables and evidence mappings. The recorded-evidence supplement contains the synthetic benchmark, frozen study configuration and protocols, all 13,824 main response records, call journal, tool projections, aggregate metrics and integrity records. Source code with its original license is supplied separately for editorial handling; the access and extraction guide explains the archive and Word representations and the numerical reproduction procedure. Model weights, provider binaries, private audit keys and credentials are not included. These materials support offline reanalysis of the recorded run; they do not independently recreate the original hardware, model generation or historical execution authorization. No public repository or public release is claimed.

REFERENCES

- Abdallah, A., Elmadany, A. A., Al Natour, S., Cavusoglu, H., Jatowt, A., & Abdul-Mageed, M. (2026). *MoCA-Agent: A market-of-claims code agent for financial and numerical reasoning* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2606.11537>
- Baldazzi, T., Bellomarini, L., Coletta, A., Iezzi, M., Maple, C., Pesare, A., & Sallinger, E. (2026). VADAOchestra: Neurosymbolic orchestration of adaptive reasoning workflows. In R. Wassermann, M.-L. Mugnier, & F. Baader (Eds.), *Proceedings of the 23rd International Conference on Principles of Knowledge Representation and Reasoning* (pp. 815–825). IJCAI Organization. <https://doi.org/10.24963/kr.2026/77>
- Berger, R. L., & Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4), 283–319. <https://doi.org/10.1214/ss/1032280304>
- Buchmann, J., Liu, X., & Gurevych, I. (2024). Attribute or abstain: large language models as long document assistants. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 8113–8140). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.463>
- Cai, X., Guo, M., Liu, J., Han, J., Guo, B., Long, Y., Chen, Y., Wu, B., Metaxas, D. N., & Li, R. (2026). *OpenPM: Auditable point-in-time evaluation for LLM portfolio-management agents* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2608.09988>
- Chen, H., Tang, X., Liu, Q., Shi, W., Li, S., Lyu, F., Luo, W., Du, X., & He, X. (2026). Fighting numerical hallucinations via data-centric compilation for online financial QA. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (pp. 7082–7093). Association for Computing Machinery. <https://doi.org/10.1145/3770855.3818407>
- Chen, W., Ma, X., Wang, X., & Cohen, W. W. (2023). Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=YfZ4ZPt8zd>
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511802843>
- Dong, Y., Zhao, E., & Chen, E. (2026). *Gold-guided programmatic distillation for financial reasoning over hybrid tables and text* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2607.14709>
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26. <https://doi.org/10.1214/aos/1176344552>
- Faria, F. T. J., Moin, M. B., Mahmud, J. A., Mridha, M. F., & Hossain, M. A. (2026). *CLAIR-Fin: An adversarial multi-agent framework for claim-level verification and adaptive debate in cross-modal financial QA* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2608.13706>
- Ferguson, N., Pennington, J., Beghian, N., Mohan, A., Kiela, D., Agrawal, S., & Nguyen, T. H. (2026). *ExtractBench: A benchmark and evaluation methodology for complex structured extraction* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2602.12247>
- Franzmeyer, T., Sravankumar, A., Liu, L., Mao, Y., Hou, R., Wang, S., Foerster, J. N., Zettlemoyer, L., & Khabsa, M. (2026). High accuracy, less talk (HALT): reliable LLMs through capability-aligned finetuning. In *The Fourteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=LYqBnNVaXD>
- Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., & Neubig, G. (2023). PAL: program aided language models. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 10764–10799). PMLR. <https://proceedings.mlr.press/v202/gao23f.html>
- Ghawate, P. (2026). *FinRCA-Bench: Benchmarking evidence retrieval and reasoning for financial AI systems* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2608.18534>
- Gu, F., Ren, X., Jiang, Z., Zhang, Z., García-Fernández, Á. F., Stefanidis, A., Zhou, M., Li, H., & Su, J. (2026). *EvidenceLens: A claim-evidence matrix for auditing financial question answering* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2606.23724>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. <https://www.jstor.org/stable/4615733>
- Hu, M., Hu, S., Wang, B., Sa, Y., Wang, X., Guo, X., Zha, D., & Xiao, J. (2026). *ECPO: Evidence-coupled policy optimization for evidence-certified candidate ranking* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2605.21993>
- Huang, T., Xu, S., Dang, J. T., Yan, S., & Yin, K. (2026). *Answer only as precisely as justified: Calibrated claim-level specificity control for agentic systems* [Preprint]. arXiv.

- <https://doi.org/10.48550/arXiv.2604.17487>
- Kan, S. (2026). *Claim-selective certification for high-risk medical retrieval-augmented generation* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2605.21949>
- Khatchadourian, R. (2026). *Replayable financial agents: A determinism-faithfulness assurance harness for tool-using LLM agents* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2601.15322>
- Kim, H. J., Kim, Y., Lee, S.-G., & Kim, T. (2025). When to speak, when to abstain: contrastive decoding with abstention. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 9710–9730). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.acl-long.479>
- Lee, Y., Kim, S., Kwak, Y., & Choo, J. (2026). BankMathBench: a benchmark for numerical reasoning in banking scenarios. In S. Piperidis, N. Bel, H. van den Heuvel, N. Ide, S. Krek, & A. Toral (Eds.), *Proceedings of the Fifteenth Language Resources and Evaluation Conference* (pp. 11010–11027). ELRA Language Resource Association. <https://doi.org/10.63317/3uxnd7yxmsb>
- Lu, J., Wang, K., Wang, Y., Tang, Q., Zeng, H., Chen, X., Pi, J., Deng, S., Chen, L., Fu, Y., Yang, K., & Sun, X. (2026). *FinToolBench: Evaluating LLM agents for real-world financial tool use* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2603.08262>
- Min, B., Edemacu, K., Cho, S.-H., Choi, Y., Jang, B., & Kim, J. W. (2026). *When absence is evidence: Evaluating completeness-sensitive negative reasoning in large language models* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2608.04591>
- 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>
- Ogunsusi, K. L., Ayisi, R. K. K., Offei, A. N. A., Giami, K. L., & Ogunsusi, K. B. (2026). Generative AI and advanced analytics for financial modeling, valuation and strategic decision-making. *American Journal of Applied Statistics and Economics*, 5(1), 97–112. <https://doi.org/10.54536/ajase.v5i1.6986>
- Parekh, K., Tiwari, A. K., & Saxena, D. (2026). *CIFQA: A deterministic tool-grounded multi-agent LLM framework for financial query answering* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2608.26114>
- Usman, R. M. (2026). *PhantomFill: When the form demands an answer, language models invent one* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2607.20492>
- Wang, Q. (2026). *AIWARE-FX: An auditable knowledge-guided AI system for measuring corporate foreign-exchange hedging disclosure* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2607.27611>
- Wang, Y., Ai, X., Patel, J., Peng, X., Mo, F., Cao, Y., Li, H., Cao, M., Qian, L., & Gutiérrez-Basulto, V. (2026). *AUDITFLOW: Executable symbolic environments for structured financial reporting verification* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2606.03031>
- Zhong, S., Zhu, J., Xu, Q., Sun, L., Wang, Y., Sun, Q., Chen, S., & Zhang, T. (2026). *FinRiskAtlas: Decision-aligned evaluation of large language models for financial risk review* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2608.25325>