Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 75-90
DOI: https://doi.org/10.51903/jtie.v5i2.541

Numerical-Reasoning Guardrails for a Quant Research Assistant: A Compact Reproducible Benchmark Using SEC and FRED Data

Zeyi Li[1], Kai Zhan*[2], Annie Wong[3]
Email: irisnycli@gmail.com
[1]*Industrial Engineering, New York University, NY, USA*
[2]*Financial Engineering, Baruch College, NY, USA*
[3]*Computer Science, Cornell Tech, NY, USA*
*Corresponding Author

Abstract

This paper presents a compact reproducible benchmark for evaluating numerical-reasoning guardrails in a quant research assistant. The revised experiment uses a fixed 2026 source snapshot derived from the SEC 2026 Q1 Financial Statement Data Sets and FRED CSV series for VIXCLS, DGS10, DGS3MO and T10Y3M. The benchmark contains 460 tasks: 300 SEC financial-ratio tasks over 50 issuer-period records, 120 FRED VIX and Treasury-rate change tasks, and 40 macro-regime classification tasks. Each answer is evaluated by five programmatic guardrails: numeric consistency, unit correctness, time-window correctness, formula correctness and citation/source consistency. Four controlled response profiles are tested: Naive-RAG, Calculator-Only, Prompted-Checklist and Guarded-Quant. These profiles are deterministic failure-mode controls rather than performance claims about any particular deployed LLM. The empirical results show that arithmetic alone is not sufficient for financial safety: Calculator-Only reaches 79.78% numeric accuracy but only 0.43% all-guardrails pass rate because source, unit, formula and window fields often fail. Guarded-Quant achieves an 88.48% all-guardrails pass rate, 97.17% numeric accuracy, 100.00% unit pass rate, 96.30% window pass rate, 98.26% formula pass rate and 96.30% citation pass rate. The findings support a modest claim: a compact benchmark can make numerical audit failures visible, but it should not be read as evidence of broad quant-assistant reliability without broader data, live model outputs and operational stress tests.

Keywords: *AI safety; Financial large language models; Guardrails; Numerical reasoning; Quant research assistant.*

I. INTRODUCTION

Quantitative research assistants increasingly summarize filings, compute ratios, compare market indicators and propose macro narratives from mixed data sources. The technical risk is not limited to hallucinated prose. A numerically fluent answer can still be unsafe when it uses a wrong denominator, mixes percentage points with basis points, shifts a time window by one observation, or omits the source needed for audit. Financial model risk management has long emphasized traceability and validation, and the same principle applies to LLM-mediated research workflows (Glasserman, 2004; Hull, 2012; Sculley et al., 2015).

This paper builds a compact benchmark for that problem. The central research question is: how much does a programmatic guardrail layer improve a quant assistant's ability to answer financial-ratio, volatility-change, rate-spread and macro-regime questions with numerically consistent and auditable outputs? The benchmark treats an answer as a structured claim consisting of a value or label, unit, time window, formula and citation. A response fails if any of these fields is inconsistent with the task truth, even when the numeric value is close.

Received on May 2026; Revised on May 2026; Accepted on July 2026; Published on August 2026

The benchmark design follows three literatures. Financial accounting and asset-pricing research show that ratios and market indicators are meaningful only when formulas, periods and units are defined precisely (Altman, 1968; Beaver, 1966; Fama & French, 1993). Volatility and macro-finance research highlights the difference between levels, changes, percent changes and basis-point changes (Bollerslev, 1986; Campbell & Shiller, 1988; Engle, 1982). AI evaluation work argues that model cards, datasheets, behavioral tests and audit processes are needed when model outputs affect decisions (Bender & Friedman, 2018; Gebru et al., 2021; Mitchell et al., 2019; Ribeiro et al., 2020; Raji et al., 2020).

The contribution is an executable evaluation harness rather than a narrative checklist. The experiment uses fixed source files, deterministic task construction and the same scoring engine for every profile. The scope is intentionally modest: the benchmark is a compact reproducibility test for numerical guardrails, not a comprehensive test of all financial reasoning behavior.

The benchmark also treats citation as a numerical-control problem. In a filing or market-data workflow, the citation is not decoration; it identifies the filing fact, series or source record that makes the calculation reconstructable. A correct-looking value without a source identifier cannot be safely reused in a notebook, report or investment memo. For that reason, citation pass rate is reported next to arithmetic pass rate throughout the paper.

The rest of the paper follows the requested manuscript structure. The literature review positions the benchmark in finance, structured financial reporting and AI safety evaluation. The research methods section defines the data, task families, response profiles, guardrails and reproducibility protocol. The results and findings section reports comparison tables and figures. The conclusion states how these findings should guide safer quant-assistant deployment and how the benchmark should be expanded.

II. LITERATURE REVIEW

Financial ratios are simple formulas, but their interpretation depends on exact field selection. Beaver (1966) and Altman (1968) established that liquidity, profitability and leverage ratios can support prediction and screening, while later empirical asset-pricing work showed that factor and return measures are sensitive to definitions and sample construction (Fama & French, 1993; Harvey et al., 2016). A quant assistant that reports gross margin, current ratio, liabilities-to-assets or debt-to-equity must therefore be tested for formula identity, not merely for plausible text.

Market and macro indicators introduce an additional unit problem. Volatility indexes are usually expressed in index points; Treasury yield movements are often communicated in basis points; spread series may be stored in percentage points but reported in basis points. The distinction is

material. ARCH and GARCH research made conditional volatility a formal object of measurement (Bollerslev, 1986; Engle, 1982), and macro-finance work has shown that yield-curve levels and spreads must be interpreted in their date context (Campbell & Shiller, 1988).

Prior research on statistical inference also motivates guardrails. White (1980) and Hansen (1982) formalized robust estimation concerns; Lo (2002) showed that even familiar performance ratios need careful distributional interpretation; Bailey et al. (2017) and Harvey et al. (2016) warned that research workflows can overstate findings when testing and reporting are not disciplined. A benchmark for financial LLM workflows should carry this discipline into answer generation by tying each value to a declared calculation path.

Structured financial reporting research is directly relevant because SEC facts are extracted from standardized but imperfect reporting systems. XBRL improves machine readability, but adoption and data-quality studies document taxonomy, consistency and validation challenges (Alles & Piechocki, 2012; Bonsón et al., 2009; Debreceny et al., 2010). These challenges imply that a research assistant should expose source identifiers, field choices and formula choices whenever it computes filing-based ratios.

Modern language-model research gives the benchmark its evaluation style. Transformers made large-scale sequence modeling practical (Vaswani et al., 2017), few-shot language models demonstrated flexible task behavior (Brown et al., 2020), and retrieval-augmented generation showed how external sources can be paired with generated text (Lewis et al., 2020). Finance-specific models such as BloombergGPT illustrate the demand for domain adaptation in financial language tasks (Wu et al., 2023). Yet retrieval and domain vocabulary do not by themselves guarantee numerical auditability.

AI documentation and assurance literature provides the safety framing. Data statements and datasheets require explicit dataset scope (Bender & Friedman, 2018; Gebru et al., 2021). Model cards recommend transparent reporting of intended use and limitations (Mitchell et al., 2019). Behavioral testing focuses on targeted capabilities rather than aggregate accuracy alone (Ribeiro et al., 2020). Internal algorithmic audits and technical-debt analyses argue that production systems need checks at every stage (Raji et al., 2020; Sculley et al., 2015). NIST's AI Risk Management Framework and ISO/IEC 23894 similarly emphasize governance, measurement and risk controls (International Organization for Standardization, 2023; National Institute of Standards and Technology, 2023).

This paper extends that literature by testing financial answers as multi-field claims. A correct answer is not only a number. It is a number or label plus the right unit, the right start and end dates, the right formula identifier and the right citation identifier. This choice reflects the practical

reality of quant research: a result that cannot be reconstructed is not safe enough for downstream decisions.

III. RESEARCH METHOD

The experiment uses a fixed 2026 source snapshot rather than a hand-entered mini table. The SEC block is derived from the SEC 2026 Q1 Financial Statement Data Sets, which include SUB, NUM, PRE and TAG files. The FRED block uses CSV files for VIXCLS, DGS10, DGS3MO and T10Y3M. Table 1 summarizes the source inventory, row counts, coverage and benchmark role.

Table 1. Dataset inventory and benchmark role.

Dataset block	Source identifiers	Raw records used	Date coverage	Benchmark use
SEC financial statements	SEC_2026Q1_FULL_ZIP; sub.txt; num.txt; pre.txt; tag.txt	6,169 submissions; 3,690,955 NUM rows	2026 Q1 archive; filing periods through 2026-02-28	SEC ratio tasks, formula checks and source-citation checks
FRED market-macro series	FRED_VIXCLS; FRED_DGS10; FRED_DGS3MO; FRED_T10Y3M	VIXCLS 9,499; DGS10 16,804; DGS3MO 11,674; T10Y3M 11,586	Common benchmark window 2026-01-02 to 2026-05-29	VIX changes, rate changes, spread checks and macro-regime classification
Source manifest	File names, source identifiers, extraction rules and SHA-256 prefixes	2026q1.zip: d18c01c615da; VIXCLS.csv: 2e7f9f4a7149; DGS10.csv: 87d9bf8b2bfa; DGS3MO.csv: 1d5748a7b8fc; T10Y3M.csv: b622229674aa	Fixed snapshot used for this run	Reproducible reconstruction and audit trail

The SEC extraction applies deterministic quality screens. The generator keeps 10-K and 10-Q submissions, requires USD facts without segment or coreg dimensions, matches fact dates to the filing period, and requires the fields needed for the six ratio formulas in Table 5. This screen identifies 958 eligible issuer-periods from the 2026 Q1 archive; the compact run uses 50 deduplicated issuer-period records. Table 2 reports the extraction flow, and Table 3 lists representative records used in ratio tasks.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 75-90
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Table 2. SEC extraction and quality-screen summary.

Extraction step	Count	Role in benchmark
SEC 2026 Q1 submissions in SUB	6,169	Raw archive scope
10-K and 10-Q submissions	5,234	Forms eligible for ratio task generation
Issuer-periods passing tag-quality screens	958	USD, no segment/coreg value, filing-period date match, required tags available
Issuer-periods selected for compact benchmark	50	Deterministic deduplicated issuer-period sample
SEC financial-ratio tasks	300	50 issuer-periods x 6 ratios

The FRED extraction uses the common observation window where VIXCLS, DGS10, DGS3MO and T10Y3M are all present. This window runs from 2026-01-02 through 2026-05-29 and contains 102 common daily observations. Table 4 reports the final common-date observations used in the market and macro tasks.

Table 3. Representative SEC issuer-period records used in ratio tasks. Revenue values are USD millions.

Company	Form	Period end	Revenue	Gross margin	Current ratio	Debt/ equity
Micron Technology Inc	10-Q	2026-02-28	37,503.0	67.73%	2.897	0.401
Adobe Inc.	10-Q	2026-02-28	6,398.0	89.62%	0.912	1.598
Winnebago Industries Inc	10-Q	2026-02-28	1,360.1	12.84%	2.296	0.661
Enerpac Tool Group Corp	10-Q	2026-02-28	299.0	48.49%	2.559	0.952
Millerknoll, Inc.	10-Q	2026-02-28	2,837.5	38.53%	1.645	1.912
Mccormick & Co Inc	10-Q	2026-02-28	1,873.9	37.83%	0.758	1.163
Cisco Systems, Inc.	10-Q	2026-01-31	30,232.0	65.22%	0.955	1.585
Palo Alto Networks Inc	10-Q	2026-01-31	5,068.0	73.90%	1.045	1.659
Applied Materials Inc /De	10-Q	2026-01-31	7,012.0	48.99%	2.715	0.733
Brady Corp	10-Q	2026-01-31	789.4	51.08%	2.131	0.390

The benchmark task generator creates three task families. SEC financial-ratio tasks ask for gross margin, operating margin, net margin, current ratio, liabilities-to-assets and debt-to-equity. FRED VIX/rate tasks ask for absolute VIX changes, percent VIX changes, Treasury basis-point changes, official T10Y3M spread levels and T10Y3M spread changes. Macro-regime tasks use a deterministic rule: VIX at or above 20 flags elevated volatility, and a negative T10Y3M spread flags a yield-curve warning. Combining the two conditions produces four labels: benign risk-on, elevated-volatility expansion, yield-curve warning and stress. Figure 1 shows the pipeline, and Figure 2 shows the resulting task composition.

Table 4. FRED Observations On The Final Common Benchmark Date.

Series	Date	Value	Unit	Source id
VIXCLS	2026-05-29	15.32	index points	FRED_VIXCLS
DGS10	2026-05-29	4.45	percent	FRED_DGS10

DGS3MO	2026-05-29	3.69	percent	FRED_DGS3MO
T10Y3M	2026-05-29	0.76 percent / 76.0 bps	percent and basis points	FRED_T10Y3M

Every task carries explicit metadata: expected value or label, answer type, unit, start date, end date, formula identifier, citation identifier, tolerance and decimal precision. Numeric tolerances are narrow: 0.01 percentage point for percentage ratios, 0.005 for unitless ratios, 0.1 basis point for rate and spread tasks, and 0.01 index point for VIX changes. Regime tasks require exact categorical matches.

Table 5. Formula catalog and required units.

Formula id	Formula used	Expected unit	Dataset family
gross_margin	gross_profit / revenue	%	SEC
operating_margin	operating_income / revenue	%	SEC
net_margin	net_income / revenue	%	SEC
current_ratio	current_assets / current_liabilities	ratio	SEC
liabilities_to_assets	liabilities / total_assets	ratio	SEC
debt_to_equity	liabilities / shareholders_equity	ratio	SEC
absolute_change	end value - start value	index points	FRED
percent_change	(end value - start value) / start value	%	FRED
basis_point_change	(end yield - start yield) x 100	basis points	FRED
T10Y3M_level	official FRED T10Y3M level x 100	basis points	FRED
T10Y3M_change	(end T10Y3M - start T10Y3M) x 100	basis points	FRED
macro_rule_vix20_spread0	VIX>=20 flags elevated volatility; T10Y3M<0 flags curve warning	label	FRED

Four controlled response profiles are evaluated. Naive-RAG represents a retrieval-style answer with incomplete formula and citation discipline. Calculator-Only represents an arithmetic-focused answer that frequently omits audit fields. Prompted-Checklist represents a schema-prompted answer that attempts to report all required fields. Guarded-Quant represents a validation-and-repair condition that checks the same structured fields used by the scoring engine. Table 7 defines the profiles. These profiles are not real LLM systems; they are controlled failure-mode generators designed to isolate the value of numerical guardrails.

Table 6. Benchmark task distribution.

Task family	Tasks
SEC financial ratio	300
FRED VIX/rate change	120
macro regime	40
Total	460

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 75-90
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

The guardrail engine evaluates five dimensions: numeric consistency, unit correctness, time-window correctness, formula correctness and citation correctness. Table 8 defines the pass criteria. The all-pass metric requires all five conditions to pass simultaneously. This conservative design matches financial audit practice: a value without a valid source, formula or date window is not treated as a safe answer.

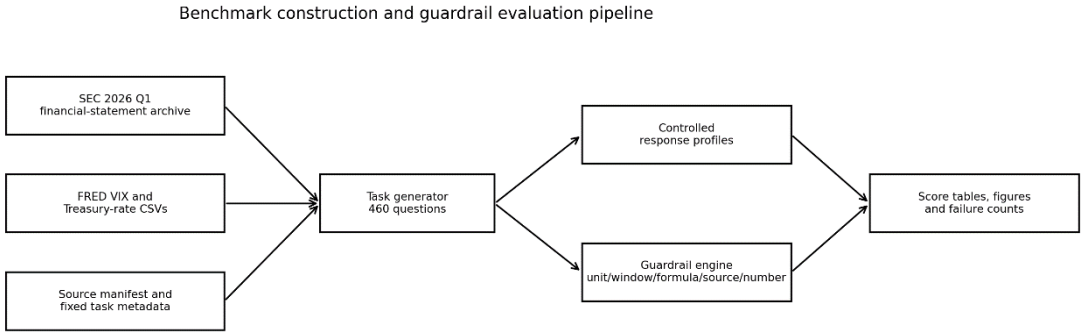

Figure 1. Benchmark construction and guardrail evaluation pipeline.

Guarded-Quant's performance should be interpreted carefully. Its validation and repair rules are intentionally aligned with the benchmark fields because the experiment asks whether enforcing those fields improves auditability. The result therefore measures the benefit of structured validation under this benchmark, not general reasoning ability across arbitrary financial tasks.

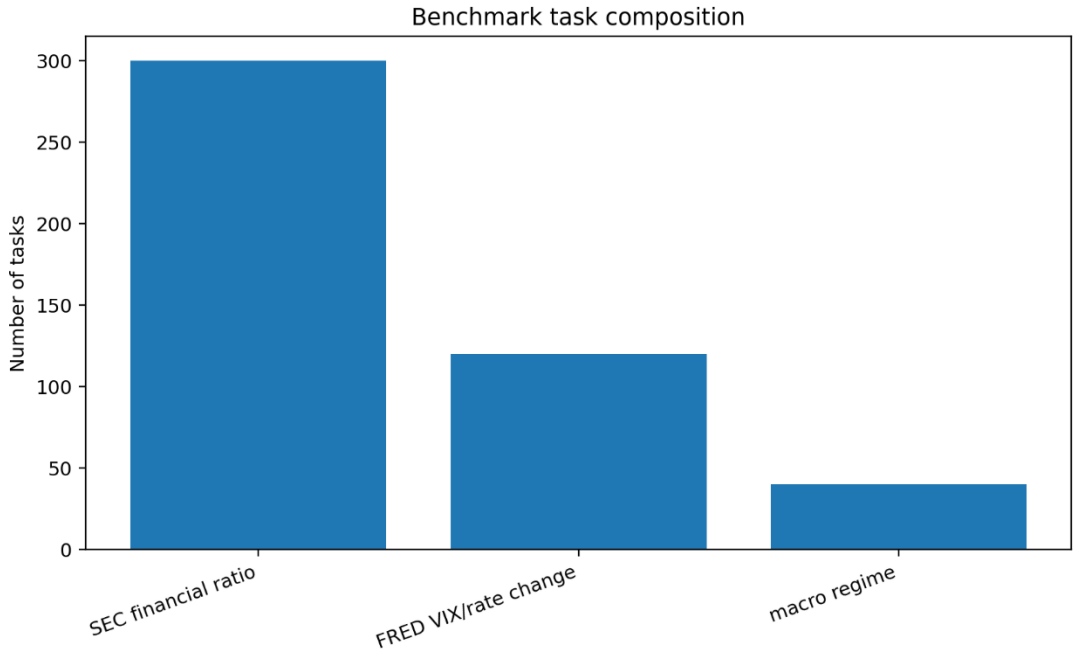

Figure 2. Task-family composition of the 460-question benchmark.

The implementation stores both rendered answer text and structured answer fields. This mirrors a practical deployment pattern in which a user sees a natural-language explanation while the application layer receives machine-checkable fields. In production, the same approach can be implemented with JSON schemas, typed function calls or constrained decoding followed by deterministic validation.

Table 7. Controlled response profiles evaluated in the experiment.

Profile	Implementation in experiment	Purpose
Naive-RAG	Controlled retrieval-style profile with weak formula and source discipline	Tests source-grounding and provenance failure modes
Calculator-Only	Controlled arithmetic-centered profile that often omits units, formulas, dates and citations	Separates numeric closeness from auditability
Prompted-Checklist	Controlled profile that attempts to report value, unit, date, formula and source fields	Measures the effect of schema prompting without deterministic repair
Guarded-Quant	Validation-and-repair profile for the same structured fields used in scoring	Measures the benefit of enforcing the benchmark audit schema

The SEC data design also clarifies limitations. The benchmark uses standard face-statement tags and excludes segment-specific, coregistrant and missing-field cases from the main task generator. That choice keeps the experiment reproducible, but it does not solve all accounting ambiguity. Restatements, alternate taxonomy tags, missing values, inconsistent source formats and data-service failures remain important stress cases for future benchmark expansions.

Table 8. Guardrail taxonomy and pass criteria.

Guardrail	Pass condition	Operational threshold
Numeric consistency	Output value or label matches benchmark truth within tolerance	0.01 percentage point for percent ratios; 0.005 for ratios; 0.1 bp for rate tasks; exact label for regime
Unit	Reported unit equals required unit	%, ratio, basis points, index points or regime label
Time window	Reported start and end dates match task metadata	Single-date tasks require start=end
Formula	Declared formula identifier matches task formula id	Prevents denominator, level/change and basis-point errors
Citation	Source identifier is present and matches task source	Supports audit trail to SEC or FRED manifest

IV. RESULT AND DISCUSSION

The experiment produces 1,840 profile-specific answers: 460 tasks multiplied by four controlled response profiles. Table 9 reports the main comparison, and Figure 3 visualizes the end-to-end all-guardrails pass rate. The central result is that numeric accuracy and financial auditability diverge. Calculator-Only reaches 79.78% numeric accuracy but only 0.43% all-pass performance because source, formula, unit and window fields often fail. Guarded-Quant reaches 88.48% all-pass performance and 97.17% numeric accuracy.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 75-90
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Table 9. Overall experimental comparison by controlled response profile.

Profile	Tasks	All-pass rate	Numeric accuracy	Unit pass	Window pass	Formula pass	Citation pass
Calculator-Only	460	0.43%	79.78%	38.48%	56.74%	44.13%	9.13%
Naive-RAG	460	11.74%	75.65%	86.74%	76.52%	46.52%	48.70%
Prompted-Checklist	460	33.70%	86.96%	90.43%	77.39%	80.87%	67.61%
Guarded-Quant	460	88.48%	97.17%	100.00%	96.30%	98.26%	96.30%

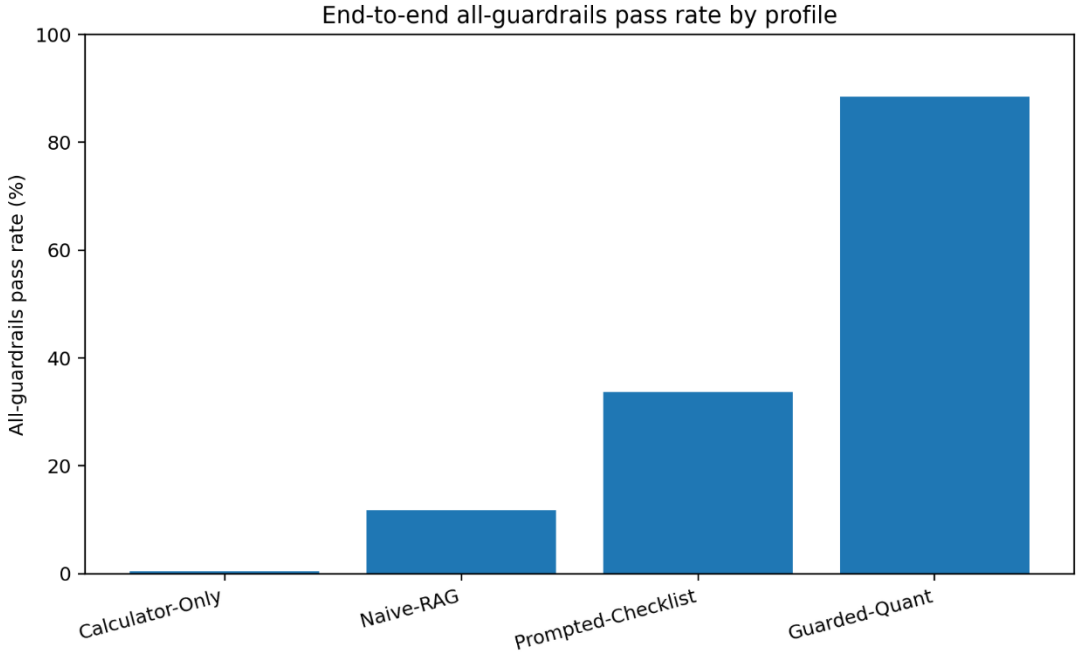

Figure 3. End-to-end all-guardrails pass rate by controlled response profile.

Table 10. Family-level all-pass and numeric-accuracy results.

Profile	Task family	All-pass rate	Numeric accuracy	Tasks
Calculator-Only	SEC financial ratio	0.33%	80.67%	300
Calculator-Only	FRED VIX/rate change	0.83%	78.33%	120
Calculator-Only	macro regime	0.00%	77.50%	40
Naive-RAG	SEC financial ratio	11.67%	77.00%	300
Naive-RAG	FRED VIX/rate change	14.17%	72.50%	120
Naive-RAG	macro regime	5.00%	75.00%	40
Prompted-Checklist	SEC financial ratio	36.67%	85.67%	300
Prompted-Checklist	FRED VIX/rate change	23.33%	88.33%	120
Prompted-Checklist	macro regime	42.50%	92.50%	40
Guarded-Quant	SEC financial ratio	88.00%	97.33%	300
Guarded-Quant	FRED VIX/rate change	90.00%	97.50%	120
Guarded-Quant	macro regime	87.50%	95.00%	40

The dimension-level results explain the difference. Figure 4 shows that Guarded-Quant reaches 100.00% unit pass rate, 96.30% window pass rate, 98.26% formula pass rate and 96.30% citation

pass rate. Calculator-Only has useful arithmetic behavior but fails 90.87% of citation checks and 61.52% of unit checks, as shown in Table 11. These results support the paper's main claim: a benchmark centered only on numerical closeness would overstate safety for finance applications.

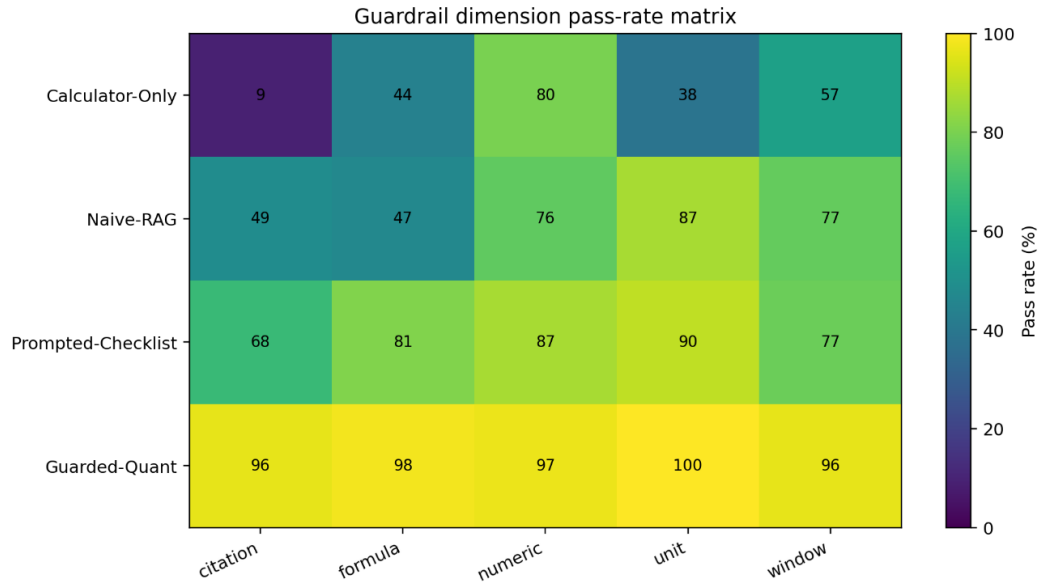

Figure 4. Guardrail dimension pass-rate matrix.

Family-level results are reported in Table 10 and Figure 5. Guarded-Quant all-pass rates are 88.00% for SEC financial-ratio tasks, 90.00% for FRED VIX/rate tasks and 87.50% for macro-regime tasks. Prompted-Checklist improves auditability relative to Naive-RAG and Calculator-Only, but it remains sensitive to citation and window errors. This pattern indicates that prompt instructions help formatting but do not reliably bind answers to exact dates, formulas and source identifiers.

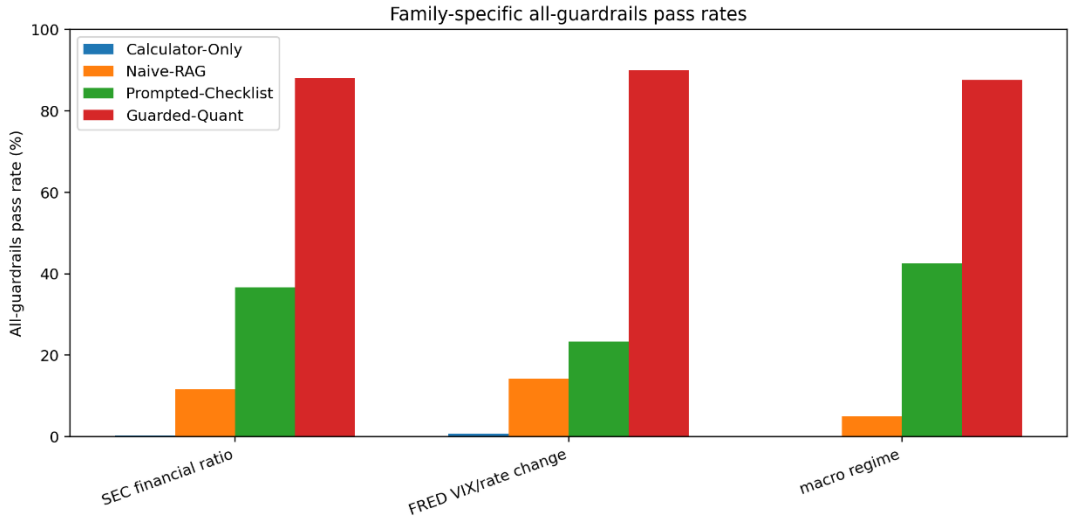

Figure 5. Family-specific all-guardrails pass rates.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 75-90
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Table 11. Guardrail failure counts by profile and dimension.

Profile	Guardrail	Failures	Failure rate
Calculator-Only	numeric	93	20.22%
Calculator-Only	unit	283	61.52%
Calculator-Only	window	199	43.26%
Calculator-Only	formula	257	55.87%
Calculator-Only	citation	418	90.87%
Naive-RAG	numeric	112	24.35%
Naive-RAG	unit	61	13.26%
Naive-RAG	window	108	23.48%
Naive-RAG	formula	246	53.48%
Naive-RAG	citation	236	51.30%
Prompted-Checklist	numeric	60	13.04%
Prompted-Checklist	unit	44	9.57%
Prompted-Checklist	window	104	22.61%
Prompted-Checklist	formula	88	19.13%
Prompted-Checklist	citation	149	32.39%
Guarded-Quant	numeric	13	2.83%
Guarded-Quant	unit	0	0.00%
Guarded-Quant	window	17	3.70%
Guarded-Quant	formula	8	1.74%
Guarded-Quant	citation	17	3.70%

Failure-count analysis identifies the remaining risks. Guarded-Quant has 13 numeric failures, zero unit failures, 17 window failures, 8 formula failures and 17 citation failures across 460 tasks. These residual failures are local and interpretable. Naive-RAG and Calculator-Only produce more diffuse failures across dimensions, which makes post-hoc correction less reliable. Table 11 summarizes the failure counts by dimension. Numeric-error distributions provide another view of the failure modes.

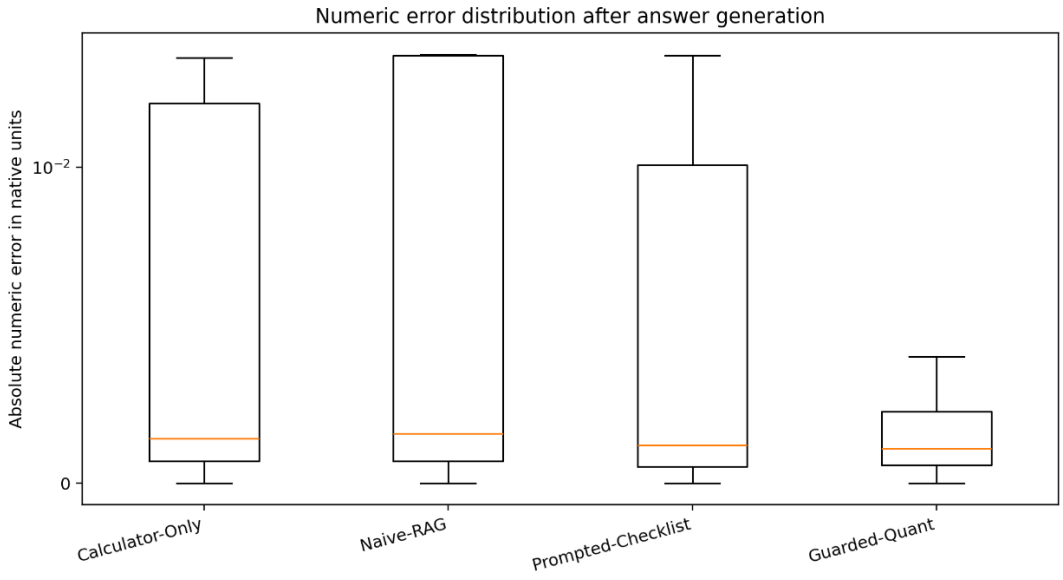

Figure 6. Numeric error distribution by controlled response profile.

Table 12 reports the median, 95th percentile and mean absolute error for numeric tasks. Figure 6 shows that the guarded profile has a much tighter central distribution, while the weaker profiles generate larger tails from decimal-versus-percent and percentage-point-versus-basis-point mistakes. A small number of unit-conversion mistakes can dominate tail risk even when median arithmetic looks acceptable.

Table 12. Numeric error summary for numeric tasks.

Profile	Median abs. error	95th pct. abs. error	Mean abs. error
Calculator-Only	0.0014	59.4989	18.2728
Naive-RAG	0.0016	81.0000	16.7930
Prompted-Checklist	0.0012	12.0751	9.4080
Guarded-Quant	0.0011	0.0222	0.7773

The macro-input figure verifies logical coherence. Figure 7 plots VIXCLS and the official T10Y3M spread during the 2026 common window. VIXCLS crosses the 20 threshold during part of the window, while the T10Y3M spread remains positive. The macro-regime tasks therefore include benign risk-on and elevated-volatility expansion cases but not yield-curve warning cases in this source snapshot.

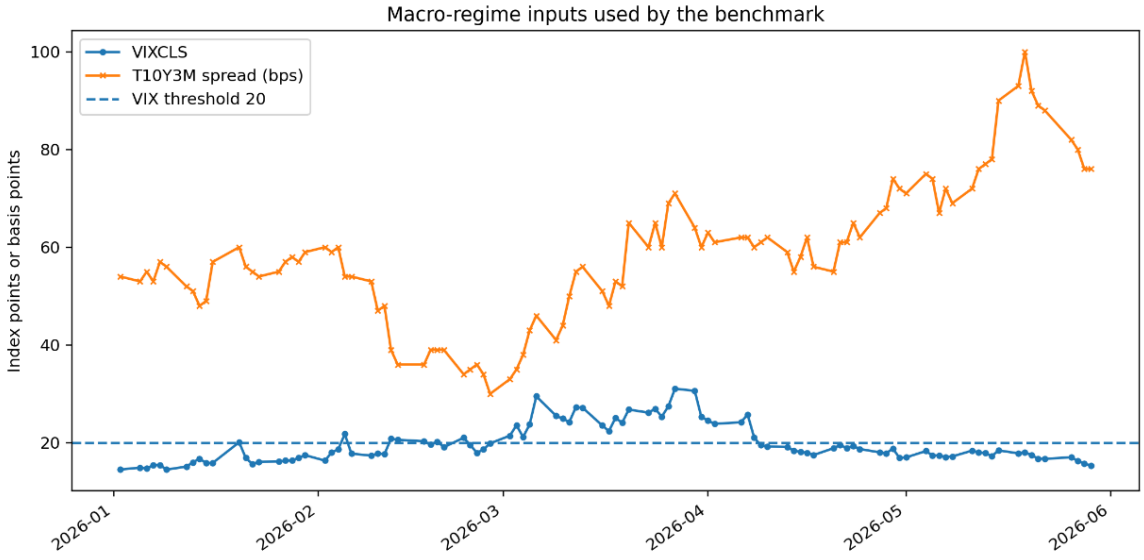

Figure 7. Macro-regime inputs used by the benchmark.

Representative records in Table 13 illustrate the specificity required. For example, a SEC gross-margin task requires the formula gross_profit / revenue, the percent unit, the filing-period date and the SEC source identifier. A T10Y3M task requires the official FRED spread series, a basis-point conversion and the correct date window. These examples show how the benchmark converts familiar quant questions into auditable structured claims.

The comparison also shows why mixed task families are useful. SEC ratio questions mainly test formula and accounting-field discipline. FRED rate questions test unit conversion and date

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 75-90
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

windows. Macro-regime questions test categorical reasoning derived from numeric thresholds. Combining the families makes the all-pass score more informative because a safe assistant must satisfy the same audit standard across different financial data workflows.

Table 13. Representative benchmark tasks and expected metadata.

Task id	Family	Question	Truth	Unit	Formula id	Citation id
SEC_723125_2026-02-28_gross_margin	SEC financial ratio	Compute gross margin for the stated SEC issuer-period.	67.731	%	gross_margin	SEC_2026Q1_3125260000006
FRED_VIXCLS_2026-01-02_2026-01-09_	FRED VIX/rate change	Compute VIXCLS change, 2026-01-02 to 2026-01-09.	-0.020	index points	absolute_change	FRED_VIXCLS
FRED_T10Y3M_2026-01-09_level	FRED VIX/rate change	Report official T10Y3M spread on 2026-01-09.	56.000	basis points	T10Y3M_level	FRED_T10Y3M
MACRO_REGIME_2026-01-02	macro regime	Classify regime on 2026-01-02 using VIX>=20 and T10Y3M<0.	benign risk-on	regime label	macro_rule_vix20_spread0	FRED_VIXCLS+FRED_T10Y3M

V. CONCLUSION AND RECOMMENDATION

This paper constructs and evaluates a compact numerical-reasoning guardrail benchmark for a quant research assistant. The benchmark uses a fixed 2026 SEC and FRED source snapshot, generates 460 tasks, evaluates 1,840 profile-specific answers and reports results from a deterministic guardrail engine. The central conclusion is that a financial assistant should not be judged by numeric accuracy alone. Financial safety requires the number or label, unit, time window, formula and source to be correct at the same time.

The best-performing configuration is Guarded-Quant, which achieves an 88.48% all-pass rate and strong dimension-level pass rates. The result demonstrates the value of combining generated answers with programmatic validation. The weaker configurations reveal the main failure modes. Calculator-Only produces many numerically close answers but fails auditability. Naive-RAG retrieves plausible source-grounded content but frequently misses formula or citation discipline. Prompted-Checklist improves behavior but does not eliminate metadata failures.

The claim should remain appropriately bounded. The experiment evaluates controlled response profiles on a compact benchmark, not live proprietary LLM systems and not the full universe of quant research tasks. Guarded-Quant is strong partly because its validation layer is designed

around the same structured fields that the benchmark scores. This is useful evidence for the value of guardrails, but it should not be overstated as evidence of broad financial reasoning ability.

The first recommendation is to deploy quant assistants with structured output schemas. Each answer should include a value or label, unit, start date, end date, formula identifier and citation identifier. The second recommendation is to evaluate all-pass accuracy rather than numeric accuracy alone. The third recommendation is to maintain a source manifest so that every cited fact can be traced back to an official filing or market-data series. The fourth recommendation is to add data-quality stress tests for restatements, ambiguous tags, missing values and source-format changes before using such a system in production.

Future work should add live LLM outputs under the same guardrail engine, expand the SEC sample beyond the compact issuer-period selection, and introduce operational stress cases such as API outages, revised observations and inconsistent source schemas. The practical implication is straightforward: a quant assistant should be treated as a calculation-and-provenance system, not merely as a chat interface that happens to discuss finance.

REFERENCES

Alles, M., & Piechocki, M. (2012). Will XBRL improve corporate governance? A framework for enhancing governance decision making using interactive data. International Journal of Accounting Information Systems, 13(2), 91-108.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589-609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2017). The probability of backtest overfitting. Journal of Computational Finance, 20(4), 39-69.

Beaver, W. H. (1966). Financial ratios as predictors of failure. Journal of Accounting Research, 4, 71-111. https://doi.org/10.2307/2490171

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587-604.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307-327. https://doi.org/10.1016/0304-4076(86)90063-1

Bonsón, E., Cortijo, V., & Escobar, T. (2009). Towards the global adoption of XBRL using International Financial Reporting Standards. International Journal of Accounting Information Systems, 10(1), 46-60.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Henighan, J., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

Campbell, J. Y., & Shiller, R. J. (1988). The dividend-price ratio and expectations of future dividends and discount factors. Review of Financial Studies, 1(3), 195-228.

Debreceny, R., Farewell, S., Piechocki, M., Felden, C., & Graning, A. (2010). Does it add up? Early evidence on the data quality of XBRL filings to the SEC. Journal of Accounting and Public Policy, 29(3), 296-306.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50(4), 987-1007.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1), 3-56.

Federal Reserve Bank of St. Louis. (2026). FRED economic data series: VIXCLS, DGS10, DGS3MO and T10Y3M. https://fred.stlouisfed.org/

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

Glasserman, P. (2004). Monte Carlo methods in financial engineering. Springer.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. Econometrica, 50(4), 1029-1054.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). ... and the cross-section of expected returns. Review of Financial Studies, 29(1), 5-68.

Hull, J. C. (2012). Risk management and financial institutions (3rd ed.). Wiley.

International Organization for Standardization. (2023). ISO/IEC 23894:2023 Artificial intelligence - Guidance on risk management. ISO.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Lewis, M., Yih, W.-T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Lo, A. W. (2002). The statistics of Sharpe ratios. Financial Analysts Journal, 58(4), 36-52.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220-229.

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 33-44.

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4902-4912.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems, 28, 2503-2511.

U.S. Securities and Exchange Commission. (2026). Financial Statement Data Sets. https://www.sec.gov/data-research/sec-markets-data/financial-statement-data-sets

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica, 48(4), 817-838.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. arXiv.