PT Jurnal Cendekia Indonesia
Journal of Information Technology and Informatics Engineering
Homepage: https://journal.jci.co.id/jitie
Vol. 02 No. 04 (2026) Page: 76-87
E-ISSN: 3123-7827

Numerical-Reasoning Guardrails for Financial Chart Vision-Language Models with Evidence Extraction, Calibrated Abstention, and Institutional Risk Dashboards

[1]Benjamin Harris, [2]Derek Guan, [3]Charlotte Martin

[1]Department of Computer Science, Stony Brook University, Stony Brook, NY, USA
[2]Department of Computer Science Oregon State University, Corvallis, OR, USA
[3]Department of Computer Science, University of Utah, Salt Lake City, UT, USA

Email: [1*]b.harris_sbu@gmail.com

Abstract

Financial chart questions require a model to align labels, periods, units, visual marks, and arithmetic before returning a compact answer. This study evaluates a backbone-agnostic numerical guardrail on the archived FinChart-Bench release containing 7,019 annotations over 1,202 chart keys. The system extracts OCR evidence, executes a restricted arithmetic trace, estimates answer correctness, and routes low-confidence cases to review. All questions were assigned to five source-document-grouped folds; each fold served once as test data, while a separate fold calibrated confidence and the remaining folds trained textual baselines. On 2,285 numerical QA records, the evidence executor reached 69.85% strict exact match and 93.39% precision-aware accuracy. Isotonic calibration reduced adaptive expected calibration error from 0.144 to 0.063 and area under the risk-coverage curve from 0.190 to 0.179. A threshold selected under a 10% calibration-fold risk budget answered 43.02% of QA records with 85.25% selective accuracy on held-out source documents. OCR recovered all operands for 30.28% of QA records and at least one operand for 68.80%, showing that arithmetic verification can remain useful even when chart text extraction is incomplete. The resulting dashboard exposes evidence completeness, calculation status, calibrated correctness, and review priority. These findings support selective, auditable assistance rather than unqualified automation in financial chart analysis.

Keywords: Financial Chart Question Answering; Vision-Language Models; Numerical Reasoning; OCR Evidence; Tool-Assisted Computation; Confidence Calibration; Selective Prediction; Institutional Risk Dashboard

Abstrak

Financial chart questions require a model to align labels, periods, units, visual marks, and arithmetic before returning a compact answer. This study evaluates a backbone-agnostic numerical guardrail on the archived FinChart-Bench release containing 7,019 annotations over 1,202 chart keys. The system extracts OCR evidence, executes a restricted arithmetic trace, estimates answer correctness, and routes low-confidence cases to review. All questions were assigned to five source-document-grouped folds; each fold served once as test data, while a separate fold calibrated confidence and the remaining folds trained textual baselines. On 2,285 numerical QA records, the evidence executor reached 69.85% strict exact match and 93.39% precision-aware accuracy. Isotonic calibration reduced adaptive expected calibration error from 0.144 to 0.063 and area under the risk-coverage curve from 0.190 to 0.179. A threshold selected under a 10% calibration-fold risk budget answered 43.02% of QA records with 85.25% selective accuracy on held-out source documents. OCR recovered all operands for 30.28% of QA records and at least one operand for 68.80%, showing that arithmetic verification can remain useful even when chart text extraction is incomplete. The resulting dashboard exposes evidence completeness, calculation status, calibrated correctness, and review priority. These findings support selective, auditable assistance rather than unqualified automation in financial chart analysis.

Kata Kunci: Financial Chart Question Answering; Vision-Language Models; Numerical Reasoning; OCR Evidence; Tool-Assisted Computation; Confidence Calibration; Selective Prediction; Institutional Risk Dashboard

A. INTRODUCTION

Financial charts compress several layers of meaning into a single visual surface. A reliable answer can depend on the identity of a series, the horizontal position of a period, the unit attached to an axis, the sign of a change, and the precision expected in the final number. A model can therefore return a fluent but unsafe response after reading the right values with the wrong scale, applying the right formula to the wrong dates, or performing correct arithmetic but formatting the result inconsistently. FinChart-Bench was created to expose these weaknesses using real financial presentation charts rather than templated graphics [1]. Its task mixture of true/false (TF), multiple choice (MC), and numerical question answering (QA) makes it suitable for studying both end answers and the controls that should surround them.

Received: 01-08-2026 | Revised: 09-08-2026 | Published Online: 10-08-2026

A useful guardrail must distinguish four failure stages. Evidence failure occurs when a label, period, or value is not recovered from the chart. Association failure occurs when recognized text is linked to the wrong bar, line, axis, or legend entry. Operation failure applies an inappropriate sum, difference, ratio, or percentage denominator. Reporting failure returns the correct underlying quantity with an incorrect unit, sign, scale, or decimal convention. End-answer accuracy collapses these stages into one bit. An evidence-first pipeline keeps them separate so that an accepted answer can carry the values and operation that produced it, while a rejected answer can be routed with a specific reason. That decomposition also makes the control layer independent of the model that proposes the initial evidence trace.

Earlier chart benchmarks established the technical progression from synthetic bar-chart recognition to logical and numerical reasoning. DVQA emphasized labels that were not fixed at training time [4], PlotQA introduced large-scale reasoning over scientific plots [3], and ChartQA combined human-authored and synthetic questions over diverse real charts [2]. ChartQAPro subsequently increased visual diversity and included questions for which abstention is a valid behavior [14]. These benchmarks show why chart understanding cannot be reduced to generic OCR: the system must associate text with marks and then transform visual relationships into an executable claim.

Chart extraction research provides two complementary foundations for this problem. ChartOCR explicitly reconstructs chart components and values [5], whereas Donut demonstrates OCR-free document parsing [9]. Pix2Struct treats screenshot parsing as a pretraining objective [8], and DePlot translates plots into a table that a language model can reason over [6]. MatCha combines chart derendering with mathematical reasoning [7]. Chart-specialized models such as UniChart [10], ChartLlama [11], ChartAssistant [12], and ChartInstruct [13] improve task adaptation through chart pretraining or instruction tuning. Their progress motivates a separate control layer: even a strong visual backbone benefits from evidence normalization, deterministic calculation, and a principled option not to answer.

Evidence-grounded retrieval provides a complementary control perspective. Evidence-chain systems have retained deterministic provenance at the claim level [29] and at the word level [30], allowing unsupported spans to be separated from otherwise fluent answers. In operational domains, private-document RAG has been paired with citation checks for troubleshooting [31], while budgeted multi-hop retrieval makes retrieval depth an explicit policy variable [32]. Bilingual incident triage extends this pattern to source-linked diagnosis and alert summarization [33], and retrieval-quality prediction treats evidence adequacy as a measurable intermediate outcome [34]. Evidence-calibrated RAG for unanswerable questions further shows why retrieval coverage and answer confidence should be estimated separately [35]. For financial charts, the analogous intermediate objects are the visible labels, values, unit, and executable operation.

Finance adds domain-specific pressure. FinQA links financial evidence to executable numerical programs [15], TAT-QA combines tables with narrative disclosures [16], and ConvFinQA tests chained numerical dialogue [17]. FinanceBench emphasizes evidence-grounded answers over enterprise financial documents [18]. More recent multimodal evaluations extend the scope to expert financial understanding [19] and challenging visual numerical reasoning [20]. In these settings, one wrong unit conversion or percentage denominator can change the meaning of an answer even when most of the explanation sounds plausible.

Recent finance-oriented work sharpens these requirements. A quant-research assistant benchmark places numerical guardrails around SEC and FRED analysis [36], while financial-search reranking combines query rewriting, hybrid retrieval, and listwise relevance control [37]. Accounting-aware retrieval for tokenized receivables links claims to disclosure evidence [38], and disclosure-review cards expose the supporting statement or note to the analyst [39]. Evidence-constrained agents extend the same principle to disclosure, settlement, and secondary-market monitoring [40]. Related credit-risk studies pair counterfactual explanation with fairness analysis [41], calibrated probability-of-default estimation with distribution-free uncertainty [42], and uncertainty-driven rejection with risk tiering [43]. A FinChart-Bench dashboard study places these controls in a visual hierarchy designed for institutional review [44]. Together, this literature motivates a chart answer that is not only numerically plausible but also traceable, calibrated, and reviewable.

Raw accuracy is consequently incomplete. Modern model confidence can be miscalibrated [21], while selective classification measures how error changes when uncertain cases are rejected [22]. SelectiveNet integrates the reject option into learning [23], and risk-control work formalizes prediction decisions under a user-selected loss constraint [24], [25]. This study uses post-hoc isotonic calibration and held-out threshold selection; it reports observed risk together with coverage rather than treating a calibration target as a guarantee.

Application-specific reliability studies also support a modular view of confidence. Uncertainty-aware late fusion calibrates modality-specific evidence before combining it [45]. Capacity forecasting for heterogeneous GPU clusters converts point forecasts into calibrated risk summaries [46], whereas selective refusal in resume-job matching treats non-answering as an explicit service behavior [47]. Conformal demand envelopes translate uncertainty into oversubscription limits [48], and trajectory-reliability prediction estimates tool-use failure without rewriting the underlying agent trace [49]. Cross-cloud forecasting emphasizes transfer under workload and topology shift [50], while regime-aware bike-demand forecasting shows how weather and season changes can alter predictive

reliability [51]. These examples differ from chart QA, but they support the same separation between task prediction, confidence estimation, and an operational decision.

Institutional review also requires a visible record of what the system read and why it acted. Model cards [26], interactive model documentation [27], and AI service FactSheets [28] all argue for structured disclosure of evidence, performance strata, and limitations. We translate those principles into a chart-specific evidence card and review dashboard. The study asks four questions: how much deterministic execution adds beyond text or OCR baselines; which operations, units, and rounding conventions dominate residual errors; whether correctness calibration improves reliability; and how much coverage remains when low-confidence answers are routed to review. The contribution is a measured guardrail architecture, a full source-grouped evaluation of the archived benchmark, and a dashboard whose counts come directly from held-out predictions.

That separation must remain visible to human reviewers. Scientific-search evidence cards organize source support, ranking trust, and visual hierarchy [52]; capacity-dashboard briefs translate forecast uncertainty into operational design choices [53]. Privacy and data-integrity risk cards make escalation conditions explicit for tool-using agents [54], while an LLM-as-design-critic study highlights the need to align machine-generated interface feedback with human judgment [55]. Evidence-constrained incident cards similarly turn distributed log evidence into an actionable review surface [56]. The dashboard in this study adopts the same general design logic but populates every field from held-out FinChart-Bench predictions rather than from a hypothetical user study..

B. METHODS

Dataset and integrity audit. The evaluated archive was the July 23, 2025 repository snapshot (SHA-256 prefix a4fa5e55a3e35bea). Its three JSON files contained 2,384 TF, 2,350 MC, and 2,285 QA records, for 7,019 annotations. This is one record per task above the 7,016 questions documented in the original benchmark paper [1]. The archive held 1,197 TF, 1,194 MC, and 1,154 QA chart files; their union comprised 1,202 chart keys and 1,198 byte-distinct images from 662 source documents. Every JSON image reference resolved. Two MC answers stored as option text were mapped unambiguously to their corresponding labels before scoring. The separately distributed QA file was byte-identical to the archive's QA JSON.

Table 1 Integrity Audit Of The Evaluated Archive

Task	Questions	Charts	Sources	Missing	Archive SHA-256 (prefix)
TF	2384	1197	660	0	a4fa5e55a3e35bea
MC	2350	1194	659	0	a4fa5e55a3e35bea
QA	2285	1154	645	0	a4fa5e55a3e35bea
Union	7019	1202	662	0	a4fa5e55a3e35bea

Grouped evaluation. The canonical image key removed the terminal question suffix, while the source-document key used the identifier preceding the first underscore. A stratified group splitter assigned each source document to one of five folds. In rotation f, fold f was the test set, fold f+1 modulo five was the calibration set, and the remaining three folds were training data. Every annotation therefore received one out-of-fold prediction, and no chart from the same financial source crossed training, calibration, and testing within a rotation. The folds remained close in total size and task composition, as shown in Table II.

Table 2 Source-Document-Grouped Five-Fold Allocation

Fold	TF	MC	QA	All	Sources
0	479	471	458	1408	138
1	478	468	457	1403	127
2	474	469	457	1400	125
3	474	469	455	1398	123
4	479	473	458	1410	149

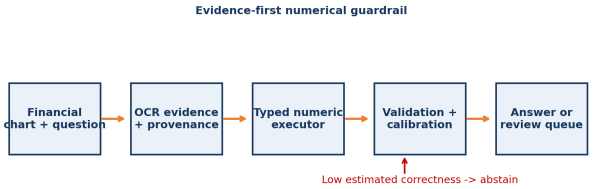

Figure 1 The OCR, Execution, Validation, Calibration, And Answer-Or-Review Flow.

OCR and image evidence. Tesseract 5.3.4 processed one representative for each byte-distinct chart with sparse-text page segmentation (PSM 11). The run retained token text, bounding boxes, token confidence, numerical tokens, image dimensions, grayscale entropy, edge strength, color variation, and dark-pixel fraction. OCR did not learn from labels, so its cache was shared across folds; downstream vectorizers, classifiers, calibrators, and thresholds were fitted inside each rotation. All 1,198 images completed successfully. Median OCR time was 0.358 s per chart and the 95th percentile was 0.865 s. Mean chart output contained 37.10 tokens and 18.14 numerical tokens.

Numerical normalization converted comma-grouped values, parenthesized negatives, percentages, basis points, and K/M/B suffixes into finite floating-point candidates while retaining the original token and location. Operand verification allowed common scale equivalents when comparing an executed number with OCR; for example, 2.7 and 2.7B remain distinguishable in the stored evidence card, but a normalized match can recognize the same chart quantity before the unit checker decides whether the answer scale is valid. This verification score was used only as a confidence feature. It never altered a reference answer or replaced the arithmetic expression selected from the evidence trace.

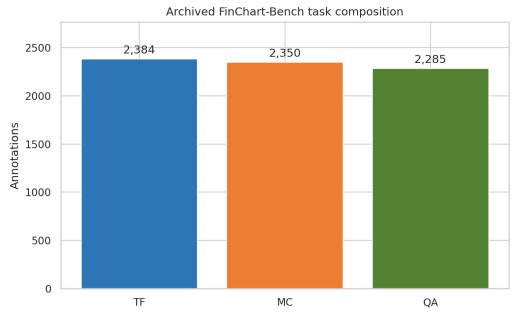

Figure 2 Question Counts In The Archived Benchmark Release.

Task representations and baselines. TF and MC used four out-of-fold systems: a training-fold majority prior; a question-only classifier; an OCR-only classifier; and question-plus-OCR fusion. Text systems combined word unigrams/bigrams with character 3-5-grams and fitted L2-regularized logistic regression with class balancing. MC input included the four choices. QA compared a training-fold median, top-1 cosine retrieval over training questions, the last numerical value in the evidence record, a stated-result parser, and a restricted arithmetic executor. Table III records the actual implementation. These compact systems establish dataset bias and guardrail behavior without attributing results to a visual backbone that was not executed.

Table 3 Systems And Evaluation Settings

System	Implementation	Inputs	Evaluation
Prior	Training-fold majority class or median	No image evidence	OOF
Question	Word/character TF-IDF + L2 logistic regression	Question and MC choices	OOF
OCR	Tesseract 5.3.4 PSM 11 + TF-IDF classifier	Chart OCR only	OOF
Question + OCR	TF-IDF fusion + L2 logistic regression	Question, choices, chart OCR	OOF
QA nearest neighbor	Question TF-IDF, cosine top-1	Training questions only	OOF
Evidence executor	Restricted AST arithmetic over evidence equation	QA evidence record + OCR verification	OOF calibration
Calibrated guardrail	Fold-local isotonic correctness calibration	Parse, OCR, complexity signals	OOF

Evidence execution. Each QA rationale was treated as an annotated evidence trace available to the guardrail evaluation. It was isolated from the TF/MC classifiers, the question-only QA retriever, and OCR text models. The parser normalized currency symbols, commas, percent signs, Unicode minus signs, and K/M/B suffixes, then located the final valid arithmetic expression preceding an equality marker; when no equality was present, it searched the last arithmetic block. An abstract-syntax-tree allowlist admitted numeric constants, parentheses, unary signs, addition, subtraction, multiplication, division, and exponentiation. Arbitrary code and nonnumeric names were rejected. Two of 2,285 traces had no executable numerical form and were marked parse failures.

Question taxonomy. A fixed priority parser assigned QA records to percentage-point difference, relative percentage change, CAGR, annualized change, average, ratio, sum, difference, extrema/range, count, or lookup/other. It separately labeled percent, basis-point, currency, count, and other units, and detected explicit years or fiscal-quarter expressions. Difference questions formed 34.5% of QA, relative percentage change 25.4%, and percentage-point difference 12.5%. Seventy-one percent contained an explicit temporal expression. Figure 3 shows the mutually exclusive operation distribution.

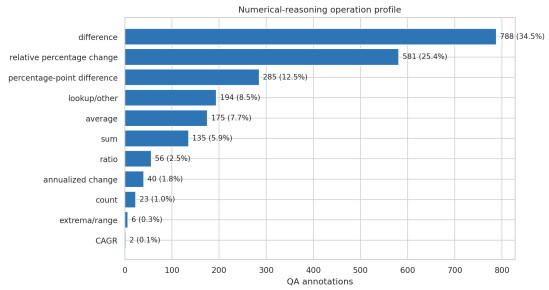

Figure 3 QA Operation Taxonomy Derived From Question Wording.

Confidence, calibration, and abstention. Raw QA confidence combined four test-label-independent signals: parser mode (weight 0.46), fraction of executed operands recovered by OCR (0.34), mean OCR confidence (0.14), and an expression-complexity term (0.06). For TF and MC, maximum class probability was the raw score. Each rotation fitted an isotonic map from raw score to empirical correctness on its calibration fold. Calibration-fold risk budgets of 5%, 10%, and 20% selected the largest confidence-ranked prefix meeting the budget; the resulting threshold was applied unchanged to the test fold. This protocol tests threshold transfer across source documents and permits the observed test risk to differ from the calibration budget.

The confidence design was therefore kept separate from answer generation. The late-fusion precedent in [45] suggests representing evidence quality explicitly before aggregation, while the capacity-control approaches in [46] and [48] illustrate how calibrated uncertainty can drive an operational threshold. Agent-trajectory reliability work [49] provides a further analogy for scoring a completed trace without changing its internal actions. These studies informed the modular structure, not the numerical weights: the four QA confidence weights in this experiment were fixed before test evaluation and were calibrated only within the grouped-fold protocol described above.

Metrics and statistics. TF and MC report accuracy, macro-F1, balanced accuracy, adaptive expected calibration error (ECE), and area under the risk-coverage curve (AURC). QA uses numeric exact match as the primary measure. A secondary precision-aware diagnostic accepts differences within max(0.05, 1% of the reference magnitude); it identifies rounding and display-precision effects but does not replace strict exact match. Evidence diagnostics include

operand OCR recall, parse coverage, and operator count. Calibration uses Brier score, negative log-likelihood, adaptive ECE, correctness AUROC, reliability plots, and AURC. Accuracy intervals used 2,000 source-document cluster bootstrap samples. Exact McNemar tests compared question-only and question-plus-OCR predictions record by record.

Run control and traceability. The experiment fixed random seed 20250720 and wrote the archive digest, library versions, fold assignment, OCR cache, out-of-fold predictions, selected thresholds, tables, and plots before manuscript assembly. Image hashes prevented repeated OCR of charts shared by TF, MC, and QA. Each saved prediction contains its record identifier, source document, fold, method, answer, confidence, and correctness outcome; QA records additionally retain the executed expression, parser mode, operand-recovery score, and answer-or-review decision. These records make aggregate values traceable to individual benchmark cases and support recomputation of every table without manual transcription.

C. RESULTS AND DISCUSSION

Dataset profile. The integrity audit found no missing images and nearly uniform fold sizes (1,398-1,410 annotations). The slight count difference from the original paper is therefore a release-version property rather than a loading error. Figure 2 and Table I preserve the task counts actually scored. Within QA, 47.0% of answers were integers, the median absolute answer was 31.25, and the 95th percentile reached 2,709.7. This wide magnitude range makes naive regression or a global numerical prior ineffective.

TF and MC baselines. Question-only TF classification achieved 71.39% accuracy, compared with 50.38% for the majority prior and 50.46% for OCR alone. The result exposes repeated linguistic direction cues such as increase, decrease, above, and below. MC behaved differently: the prior reached 41.62%, while question-only and fused models scored 40.04% and 39.91%. OCR alone reached 34.30%. Thus, MC label imbalance was stronger than the compact text signal, and unstructured OCR did not reconstruct the visual relations needed to choose among options. Adding OCR changed TF by -0.17 percentage points and MC by -0.13 points; exact McNemar p-values of 0.868 and 0.927 provide no evidence of improvement. Table IV and Figure 4 report the complete comparisons.

The contrast between TF and MC is informative for guardrail design. A question-only model can exploit the wording of a binary proposition without demonstrating that it located the referenced series, whereas MC requires the system to discriminate among four chart-dependent alternatives. The OCR fusion model supplied many correct labels and numbers but removed two-dimensional relationships when it linearized the chart. Its near-zero change over question-only prediction is therefore not evidence that the image is irrelevant; it shows that token presence alone is a weak representation of position, series membership, and visual order. A production VLM can replace this compact backbone while preserving the same evidence, execution, and abstention interfaces.

Table 4 Out-Of-Fold TF And MC Performance

Task	Method	Accuracy	95% CI	Macro-F1	ECE	AURC
MC	ocr	34.3%	32.2%-36.2%	0.268	0.046	0.627
MC	prior	41.6%	39.8%-43.7%	0.147	0.044	0.601
MC	question	40.0%	38.1%-41.9%	0.360	0.056	0.493
MC	question_ocr	39.9%	37.9%-41.9%	0.344	0.031	0.486
TF	ocr	50.5%	50.0%-51.0%	0.504	0.014	0.497
TF	prior	50.4%	49.9%-50.9%	0.335	0.008	0.498
TF	question	71.4%	69.4%-73.2%	0.714	0.037	0.175
TF	question_ocr	71.2%	69.3%-73.1%	0.712	0.042	0.191

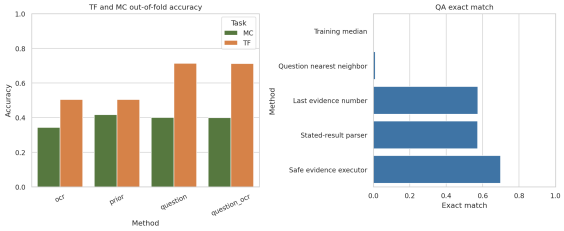

Figure 4 Out-Of-Fold Classification Accuracy And QA Exact Match.

QA predictive performance. The training-fold median never matched a held-out answer, and question nearest-neighbor retrieval produced 24 exact answers (1.05%). Evidence-based methods were substantially stronger. Taking the last numerical value yielded 57.33% strict exact match; the stated-result parser yielded 57.29%; and the safe executor reached 69.85%, or 1,596 of 2,285 questions. Its source-clustered 95% confidence interval was 67.82%-71.82%. Precision-aware accuracy was 93.39%, revealing a large gap between mathematical agreement and exact output formatting. This is not a reason to discard strict scoring: in a financial workflow, unit and precision conventions must be explicit. It does show why an audit dashboard should separate calculation failure from rounding-only mismatch.

The opening gross-margin chart provides a representative successful trace. OCR recovered 36.0% for 1Q11 and

33.4% for 1Q15, the parser selected 36.0 - 33.4, and the executor returned 2.6 percentage points. That record satisfied strict exact match and carried complete operand evidence. In contrast, the expression (128 / 252) x 100 produced 50.793651... for a reference stored as 50.79. The arithmetic and denominator were correct, but strict exact match failed until a precision convention was applied. Keeping that case in the strict-error count preserves benchmark comparability, while the precision-aware column identifies its operational character.

Together, these cases motivate a two-layer answer contract. The calculation layer retains full numerical precision, the parsed operation, and the unmodified source quantities; the reporting layer applies the unit and decimal convention expected by the question. A reviewer can then distinguish whether a disagreement began in evidence selection, arithmetic, or presentation. This distinction is particularly important for financial ratios because converting a decimal to percent is a scale decision rather than a cosmetic formatting step. The evaluated executor preserved both forms in its trace and scored the final reported value against the benchmark answer, so a numerically close output did not erase a unit or scale mismatch.

Table 5 Numerical QA Performance

Method	Strict EM	Precision-aware	95% CI	Correct answers
Training median	0.0%	0.4%	0.0%-0.0%	0
Question nearest neighbor	1.1%	1.4%	0.5%-1.7%	24
Last evidence number	57.3%	58.1%	54.8%-60.0%	1310
Stated-result parser	57.3%	58.2%	54.7%-59.9%	1309
Safe evidence executor	69.8%	93.4%	67.8%-71.8%	1596

Operation-level behavior. Difference and percentage-point questions were the most reliable, with 93.40% and 96.14% strict exact match. Sums reached 87.41% and ratios 91.07%. Relative percentage change was the principal strict-scoring failure: only 19.45% matched exactly, while 95.35% met the precision-aware criterion. The executor often computed an unrounded repeating decimal when the reference stored two decimal places. Average and annualized questions also mixed exact arithmetic with display conventions. The operation table therefore separates strict and precision-aware accuracy rather than presenting one aggregate number.

The operation strata also show why a universal numerical tolerance would be unsafe. A difference expressed in percentage points, a relative percentage change, and a ratio can contain the same operands while answering different questions. Treating all three as interchangeable after rounding would conceal a denominator error. The diagnostic tolerance was therefore used only to classify near-miss behavior after strict scoring. In deployment, the reporting policy should be chosen from the requested unit and operation, recorded in the evidence card, and applied before release. Cases without a recognized unit or with disagreement between the parsed operation and question wording should remain review candidates even when the computed number is close to a plausible answer.
Table VI.

Table 6 QA Execution And OCR Evidence By Operation

Operation	Questions	Strict EM	Precision-aware	Operand OCR	Operators
difference	788	93.4%	93.5%	57.6%	0.98
relative percentage change	581	19.4%	95.4%	40.6%	2.14
percentage-point difference	285	96.1%	96.8%	62.9%	1.07
lookup/other	194	68.6%	89.7%	36.2%	1.53
average	175	72.0%	92.6%	34.9%	1.30
sum	135	87.4%	91.9%	46.6%	1.61
ratio	56	91.1%	94.6%	60.9%	1.16
annualized change	40	57.5%	77.5%	30.0%	0.97
count	23	69.6%	69.6%	31.2%	0.52
extrema/range	6	83.3%	100.0%	36.1%	1.33
CAGR	2	50.0%	50.0%	25.0%	1.00

Evidence extraction. Across QA records, mean executed-operand OCR recall was 48.98%; at least one operand appeared in OCR for 68.80%, and every operand appeared for 30.28%. Exact answers were still possible when OCR evidence was incomplete because the benchmark evidence trace supplied normalized quantities and operations. Of the 1,596 strict-correct cases, 612 had complete OCR operand recovery and 984 had incomplete recovery. The remaining errors comprised two parse failures, three percent-scale mismatches, 538 rounding-only exact-match misses, and 146 substantive evidence-answer mismatches. The last category includes approximation differences, trace/reference conflicts, or execution choices not captured by the last-expression rule.

The 146 substantive mismatches deserve separate review from formatting misses. Some traces contain several chained equations, making the last valid expression a derived intermediate rather than the requested output. Some questions ask for an approximation read from a low-resolution visual mark, so a small evidence discrepancy changes the computed result. A smaller set contains a conflict between the narrated chart values and the stored answer. The guardrail does not silently reconcile these cases against the reference; it preserves the trace, reports failed agreement, and lets confidence features influence routing. This behavior is preferable to presenting a single aggregate accuracy that obscures annotation and evidence conflicts.

Calibration. Isotonic calibration improved every reported correctness-reliability measure. Brier score fell from 0.211 to 0.192, adaptive ECE from 0.144 to 0.063, negative log-

likelihood from 0.610 to 0.605, and AURC from 0.190 to 0.179. Correctness-detection AUROC increased from 0.639 to 0.679. Figure 5 shows that calibration removed the largest oscillations in empirical accuracy, although several confidence regions remained above or below the ideal diagonal. This residual structure is expected when the score depends on OCR completeness and expression form but not on semantic agreement between chart, rationale, and reference.

The improvement has two practical interpretations. Lower Brier score and ECE make a displayed correctness estimate more faithful on average, while lower AURC improves the ordering used to construct a review queue. Neither property changes the executed answer itself. The fold-local protocol also prevents a convenient but misleading fit on the test outcomes: each isotonic map was learned from a different source-document fold and then applied to previously unseen sources. Residual miscalibration therefore captures genuine transfer friction across issuers, chart styles, and reporting periods. The reliability diagram should be read as an operating diagnostic that accompanies the risk-coverage curve, not as evidence that one probability is valid for every future chart population.

The residual source-fold shift is consistent with other application settings in which a useful ranking score can remain imperfectly calibrated after deployment conditions change. Evidence-calibrated RAG reports separate evidence-sufficiency and answerability behavior [35]; model-risk probability-of-default workflows retain calibration and uncertainty diagnostics beside discrimination metrics [42]; and selective matching systems report refusal behavior together with accepted-case accuracy [47]. Regime-aware demand forecasting likewise treats changing exogenous conditions as a reason to re-estimate uncertainty rather than reuse a static threshold [51]. The present results add a chart-specific instance: isotonic calibration improved ordering and average reliability, yet the selected risk budget still required monitoring on new source documents.

Table 7 QA Correctness Calibration

Confidence	Brier	Adaptive ECE	NLL	Error-detection AUROC	AURC
Raw	0.211	0.144	0.610	0.639	0.190
Isotonic	0.192	0.063	0.605	0.679	0.179

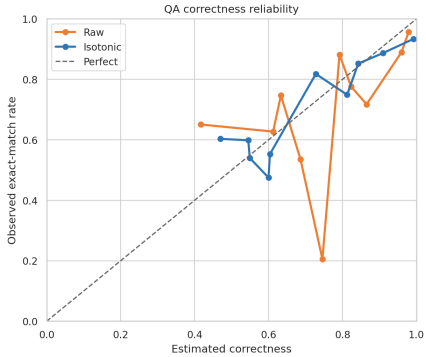

Figure 5 Raw And Isotonic QA Correctness Reliability.

Selective prediction. Ordering by calibrated correctness reduced error as coverage fell. At fixed coverage of 40%, selective risk was 14.88%, compared with 30.15% at full coverage; 60% and 80% coverage produced risks of 20.86% and 27.68%. Calibration-budget thresholds did not transfer perfectly. The 10% calibration-fold budget accepted 983 records (43.02%) and achieved 85.25% accuracy, corresponding to 14.75% observed risk. The 5% budget accepted 27.44% at 12.44% risk, while the 20% budget accepted 96.67% at 29.65% risk. These gaps are operationally important: a held-out calibration target should be displayed as a selection rule, not a promise about future error.

Expressed per 100 numerical questions, the 10% calibration-budget rule automatically answered about 43 and routed 57 to review. Roughly 37 of the 100 were accepted correctly, six were accepted incorrectly, 33 were correct but still reviewed, and 24 were both incorrect and reviewed. The false-review share is the price of using a compact, monotone confidence function. It may be reduced by a stronger evidence extractor or richer semantic consistency features, but it cannot be hidden when reporting selective accuracy. Coverage and error must travel together in any institutional summary.

The gap between a 10% selection budget and 14.75% observed held-out risk illustrates the governance problem directly. A threshold is an operating choice learned under one calibration distribution; it is not a permanent risk limit. An institution can still use the rule, but it should pair the chosen threshold with the calibration sample, date, archive digest, achieved coverage, and subsequent realized error. A change in document source or rendering style can then trigger recalibration before silent degradation becomes material. The same record also supports capacity planning: the measured 43.02% coverage specifies the automated share, while the remaining 56.98% defines the review load that staffing and service-level targets must absorb.

Deployment and operational monitoring extend the same separation principle. Hybrid-cloud LLM architecture emphasizes versioned routing and cost-aware placement [57], while evidence-based self-verification checks a candidate response before release [58]. Retrieved normal prototypes attach a concrete comparator to anomaly explanations [59], and behavior-level refusal separates unsafe action from useful response content [60]. Conformal alert control turns anomaly scores into auditable incident tickets [61], whereas continual red-teaming motivates updating a guardrail as failure patterns drift [62]. Multimodal tracking demonstrates modular evidence fusion [63], retrieval-summary integration separates finding from explaining [64], and window-level intrusion localization preserves evidence for forensic narratives [65]. These are architectural analogies rather than empirical evidence about FinChart-Bench; their relevance is the shared pattern of evidence retention, secondary reliability assessment, and explicit routing.

Table 8 Held-Out Operating Points Selected By Calibration-Fold Risk Budgets

Budget	Coverage	Observed risk	Accuracy	Accepted	Abstained
5.0%	27.4%	12.4%	87.6%	627	1658
10.0%	43.0%	14.8%	85.2%	983	1302
20.0%	96.7%	29.7%	70.3%	2209	76

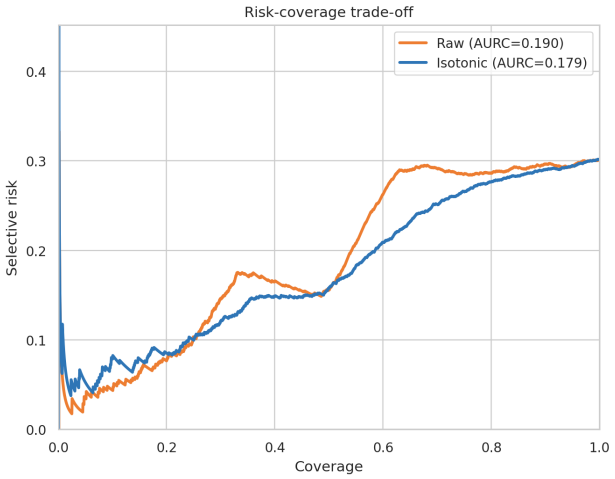

Figure 6 QA Risk-Coverage Curves Before And After Calibration.

Ablation. Parsing only the last number or stated result produced similar strict accuracy but poor ranking. The full executor raised exact match by 12.52 points over last-number extraction. Adding OCR evidence to parser confidence did not change answers, but reduced Brier score from 0.276 to 0.211 and AURC from 0.283 to 0.190. Isotonic calibration then reduced ECE from 0.144 to 0.063 and AURC to 0.179. The distinction matters: execution determines answer quality, while OCR evidence and calibration determine whether the system recognizes which answers deserve review.

Table 9 Guardrail Component Ablation

Variant	Strict EM	Brier	ECE	AURC
Last-number extraction	57.3%	0.294	0.223	0.428
Stated-result extraction	57.3%	0.245	0.023	0.430
Executor; parse confidence only	69.8%	0.276	0.236	0.283
Executor + OCR evidence	69.8%	0.211	0.144	0.190
Executor + OCR + calibration	69.8%	0.192	0.063	0.179

Dashboard interpretation. Figure 7 converts the 10% calibration-budget operating point into an audit queue. Of all QA records, 36.7% were accepted and correct, 6.3% were accepted and wrong, 33.2% were abstained despite being correct, and 23.8% were abstained and wrong. Relative percentage change generated the highest review rate, consistent with its formatting sensitivity; percentage-point differences and ordinary differences required less review. The dashboard is therefore a routing and evidence surface. It does not establish analyst productivity, financial benefit, or autonomous decision quality, because those outcomes were not measured.

Each dashboard outcome is backed by a case-level evidence card. It identifies the chart and source, shows the question and parsed operation, lists normalized operands with OCR recovery, records raw and calibrated correctness, and states the routing threshold. This gives a reviewer a concrete starting point and creates a compact evidence trail for released answers. Aggregating the same fields by operation and source separates recurring weaknesses from isolated cases, directing remediation toward extraction, calculation, reporting policy, or calibration.

The review surface is consequently designed as an evidence interface rather than a decorative result summary. Its source-and-calculation block parallels accounting disclosure cards [39] and scientific-search evidence cards [52]; its confidence and queue fields follow the risk-communication logic of financial and capacity dashboards [44], [53]. The explicit review reason is consistent with privacy and data-integrity risk cards [54] and incident visualization cards [56], while the separation between machine critique and human judgment echoes the design-critic findings in [55]. This organization does not claim that the dashboard has been usability-tested; it specifies what an institutional reviewer can inspect and how each display element maps back to a saved prediction.

Table 10 Mutually Exclusive QA Outcome Taxonomy

Error category	Questions	Share
Parse failure	2	0.1%
Percent-scale mismatch	3	0.1%
Rounding-only exact-match miss	538	23.5%
Substantive evidence-answer mismatch	146	6.4%
Correct with incomplete OCR operands	984	43.1%
Correct with complete OCR operands	612	26.8%

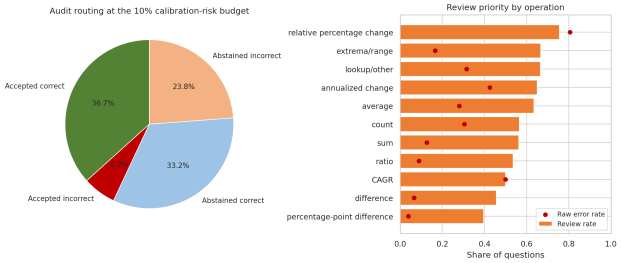

Figure 7 Measured Answer-Versus-Review Routing And Operation-Level Review Priority.

Scope and limitations. FinChart-Bench charts originate from 2015-2024 financial presentations [1], and many visual values are approximate. QA alone includes evidence traces; TF and MC do not supply equivalent rationales. The archive has no pixel-level gold evidence regions, so this study reports operand recovery rather than localization accuracy. Tesseract text plus linear models are transparent controls, not substitutes for a chart-specialized VLM. The executor measures the value of evidence normalization and deterministic arithmetic once an evidence trace exists; it does not estimate end-to-end image-to-trace accuracy. Finally, calibration shifted across source folds, and the risk-budget experiments show that thresholds require monitoring when chart sources, periods, or rendering styles change.

Cross-domain guardrails should not be mistaken for direct validation. The hybrid-cloud, anomaly-prototype, and conformal-ticket studies [57], [59], [61] address deployment or incident workflows rather than financial chart arithmetic. Likewise, self-verification, staged refusal, and continual red-teaming [58], [60], [62] target adversarial or behavioral risk, while fusion, code-intelligence, and host-intrusion studies [63]-[65] concern different evidence modalities. They are included to position the control architecture within a broader defense-in-depth literature; all quantitative claims in this article remain confined to the evaluated FinChart-Bench archive.

D. CONCLUSIONS

The full archived FinChart-Bench evaluation answers the four research questions with measured trade-offs. First, deterministic evidence execution materially outperformed numerical priors and question retrieval, reaching 69.85% strict exact match and 93.39% precision-aware accuracy on 2,285 QA records. Second, ordinary differences, percentage-point differences, sums, and ratios were handled well, while relative percentage change exposed the benchmark's strongest rounding and output-precision sensitivity. OCR recovered at least one executed operand in 68.80% of cases, but complete operand recovery in only 30.28%, confirming that evidence quality must remain visible. Third, fold-local isotonic calibration reduced ECE from 0.144 to 0.063 and AURC from 0.190 to 0.179. Fourth, abstention reduced error at the cost of substantial review: a 10% calibration-fold budget answered 43.02% of cases with 85.25% held-out accuracy. Because observed risk exceeded the calibration budget, the dashboard reports both quantities rather than treating the threshold as a guarantee. The resulting design separates visual evidence, executable arithmetic, correctness estimation, and human review. It is best used as a post-model control layer for auditable financial analysis, with strict unit, precision, and provenance checks maintained alongside any future VLM backbone.

Taken together with evidence-chain and financial guardrail work [29], [36], [44], the results support a layered design in which evidence, execution, calibration, and review remain separately inspectable.

E. REFERENCES

[1] D. Shu, H. Yuan, Y. Wang, Y. Liu, H. Zhang, and M. Du, "FinChart-Bench: Benchmarking financial chart comprehension in vision-language models," in Proc. 64th Annu. Meeting Assoc. Comput. Linguistics, 2026, pp. 13447-13466, doi: 10.18653/v1/2026.acl-long.615.

[2] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, "ChartQA: A benchmark for question answering about charts with visual and logical reasoning," in Findings ACL, 2022, pp. 2263-2279, doi: 10.18653/v1/2022.findings-acl.177.

[3] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar, "PlotQA: Reasoning over scientific plots," in Proc. IEEE WACV, 2020, pp. 1527-1536, doi: 10.1109/WACV45572.2020.9093523.

[4] K. Kafle, B. Price, S. Cohen, and C. Kanan, "DVQA: Understanding data visualizations via question answering," in Proc. IEEE/CVF CVPR, 2018, pp. 5648-5656, doi: 10.1109/CVPR.2018.00592.

[5] J. Luo, Z. Li, J. Wang, and C.-Y. Lin, "ChartOCR: Data extraction from charts images via a deep hybrid framework," in Proc. IEEE/CVF WACV, 2021, pp. 1916-1924, doi: 10.1109/WACV48630.2021.00196.

[6] F. Liu et al., "DePlot: One-shot visual language reasoning by plot-to-table translation," in Findings ACL, 2023, pp. 10381-10399, doi: 10.18653/v1/2023.findings-acl.660.

[7] F. Liu et al., "MatCha: Enhancing visual language pretraining with math reasoning and chart derendering," in Proc. ACL, 2023, pp. 12756-12770, doi: 10.18653/v1/2023.acl-long.714.

[8] K. Lee et al., "Pix2Struct: Screenshot parsing as pretraining for visual language understanding," in Proc. ICML, vol. 202, 2023, pp. 18893-18912.

[9] G. Kim et al., "OCR-free document understanding transformer," in Computer Vision-ECCV 2022, vol. 13688, pp. 498-517, doi: 10.1007/978-3-031-19815-1_29.

[10] A. Masry, P. Kavehzadeh, X. L. Do, E. Hoque, and S. Joty, "UniChart: A universal vision-language pretrained model for chart comprehension and reasoning," in Proc. EMNLP, 2023, pp. 14662-14684, doi: 10.18653/v1/2023.emnlp-main.906.

[11] Y. Han et al., "ChartLlama: A multimodal LLM for chart understanding and generation," arXiv:2311.16483, 2023, doi: 10.48550/arXiv.2311.16483.

[12] F. Meng et al., "ChartAssistant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning," in Findings ACL, 2024, pp. 7775-7803, doi: 10.18653/v1/2024.findings-acl.463.

[13] A. Masry, M. Shahmohammadi, M. R. Parvez, E. Hoque, and S. Joty, "ChartInstruct: Instruction tuning for chart comprehension and reasoning," in Findings ACL, 2024, pp. 10387-10409, doi: 10.18653/v1/2024.findings-acl.619.

[14] A. Masry et al., "ChartQAPro: A more diverse and challenging benchmark for chart question answering," in Findings ACL, 2025, pp. 19123-19151, doi: 10.18653/v1/2025.findings-acl.978.

[15] Z. Chen et al., "FinQA: A dataset of numerical reasoning over financial data," in Proc. EMNLP, 2021, pp. 3697-3711, doi: 10.18653/v1/2021.emnlp-main.300.

[16] F. Zhu et al., "TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance," in Proc. ACL-IJCNLP, 2021, pp. 3277-3287, doi: 10.18653/v1/2021.acl-long.254.

[17] Z. Chen, S. Li, C. Smiley, Z. Ma, S. Shah, and W. Y. Wang, "ConvFinQA: Exploring the chain of numerical reasoning in conversational finance

question answering," in Proc. EMNLP, 2022, pp. 6279-6292, doi: 10.18653/v1/2022.emnlp-main.421.

[18] P. Islam, A. Kannappan, D. Kiela, R. Qian, N. Scherrer, and B. Vidgen, "FinanceBench: A new benchmark for financial question answering," arXiv:2311.11944, 2023, doi: 10.48550/arXiv.2311.11944.

[19] Z. Gan et al., "MME-Finance: A multimodal finance benchmark for expert-level understanding and reasoning," in Proc. ACM Multimedia, 2025, pp. 12867-12874, doi: 10.1145/3746027.3758230.

[20] Z. Tang et al., "FinMMR: Make financial numerical reasoning more multimodal, comprehensive, and challenging," in Proc. IEEE/CVF ICCV, 2025, pp. 3245-3257.

[21] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. ICML, vol. 70, 2017, pp. 1321-1330.

[22] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in Advances in Neural Information Processing Systems 30, 2017.

[23] Y. Geifman and R. El-Yaniv, "SelectiveNet: A deep neural network with an integrated reject option," in Proc. ICML, vol. 97, 2019, pp. 2151-2159.

[24] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal risk control," in Proc. ICLR, 2024.

[25] S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. I. Jordan, "Distribution-free, risk-controlling prediction sets," J. ACM, vol. 68, no. 6, Art. 43, pp. 1-34, 2021, doi: 10.1145/3478535.

[26] M. Mitchell et al., "Model cards for model reporting," in Proc. FAT*, 2019, pp. 220-229, doi: 10.1145/3287560.3287596.

[27] A. Crisan, M. Drouhard, J. Vig, and N. Rajani, "Interactive model cards: A human-centered approach to model documentation," in Proc. ACM FAccT, 2022, pp. 427-439, doi: 10.1145/3531146.3533108.

[28] M. Arnold et al., "FactSheets: Increasing trust in AI services through supplier's declarations of conformity," IBM J. Res. Develop., vol. 63, no. 4/5, pp. 6:1-6:13, 2019, doi: 10.1147/JRD.2019.2942288.

[29] J. Jin, "Evidence-Chain Reliable RAG: Hallucination Detection, Source Attribution, and Deterministic Provenance Explanations," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 520-533, Aug. 2025, doi: 10.51903/jtie.v4i2.535.

[30] C. Li, J. Bai, and S. Wang, "Evidence-Chain Reliable RAG: Word-Level Hallucination Detection, Source Attribution, and Provenance Explanation for LLM Applications," J. Adv. Comput. Syst., vol. 4, no. 2, pp. 76-92, Feb. 2024, doi: 10.69987/JACS.2024.40207.

[31] B. Zhang, H. Rao, and D. Zhao, "Evidence-Grounded RAG for Cloud-Native DevOps: Hallucination-Resistant AIOps Question Answering over Private Operations Documents," J. Adv. Comput. Syst., vol. 4, no. 3, pp. 109-125, Mar. 2024, doi: 10.69987/JACS.2024.40308.

[32] W. Su, S. Chen, and C. Zhao, "Budgeted Multi-Hop Retrieval Agent for Compositional Question Answering: A Retrieval-Policy Evaluation on the Official MultiHop-RAG Benchmark," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 649-662, Dec. 2025, doi: 10.51903/jtie.v4i3.543.

[33] G. Liu, C. Li, and E. Zhang, "OpsLLM for Cloud Incident Triage: Bilingual RAG-Based Root Cause Analysis and Alert Summarization for AI Infrastructure Operations," J. Adv. Comput. Syst., vol. 4, no. 4, pp. 97-111, Apr. 2024, doi: 10.69987/JACS.2024.40408.

[34] R. Zhang, Z. Wen, C. Wang, C. Tang, P. Xu, and Y. Jiang, "Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms," arXiv:2511.19481, 2025, doi: 10.48550/arXiv.2511.19481.

[35] Z. S. Zhong, J. Chen, E. Zhong, and X. Sun, "Evidence-Calibrated RAG for Unanswerable Question Answering: Retrieval Coverage, Abstention Calibration, and Hallucination-Proxy Analysis on SQuAD 2.0," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 502-520, Aug. 2025, doi: 10.51903/jtie.v4i2.536.

[36] Z. Li, K. Zhang, and A. Wong, "Numerical-Reasoning Guardrails for a Quant Research Assistant: A Compact Reproducible Benchmark Using SEC and FRED Data," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 75-90, Jun. 2026, doi: 10.51903/jtie.v5i2.541.

[37] S. Meng, J. Chen, and I. Zheng, "LLM-Inspired Offline Reranking for Financial Search: Query Rewriting, Hybrid Retrieval, and Listwise Relevance Ranking on FiQA," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 361-378, Apr. 2026, doi: 10.51903/jtie.v5i1.537.

[38] Y. Chen, S. Zhou, and E. Lin, "Accounting-Aware Evidence Retrieval for Institutional Due Diligence of Tokenized Trade Receivable RWA," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 649-663, Dec. 2025, doi: 10.51903/jtie.v4i3.542.

[39] K. Zhang, Y. Chen, and A. Qian, "Evidence-Grounded Accounting Disclosure Review Cards: A Visual Communication Framework for LLM-Style Explanations over SEC Financial Statements and Notes," Int. J. Graph. Des., vol. 3, no. 2, p. 395, Oct. 2025, doi: 10.51903/ijgd.v3i2.3710.

[40] S. Zhou, Y. Chen, and K. Lee, "Accounting-Aware Evidence-Constrained Agents for Disclosure, Settlement, and Secondary-Market Risk Monitoring in Tokenized," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 60-74, 2026, doi: 10.51903/jtie.v5i2.544.

[41] Q. Xin, "Explainable and Fair Credit Risk Scoring with Counterfactual Explanations: A Reproducible Evaluation on the German Credit Dataset (HELOC-Motivated)," J. Inf. Technol., vol. 14, no. 2, pp. 215-231, Jun. 2026, doi: 10.32664/j-intech.v14i02.2228.

[42] J. Jin, T. Huang, and S. Lu, "A Model-Risk-Friendly Probability of Default Workflow: Calibration, Distribution-Free Uncertainty Quantification, and SHAP Explanations on the UCI Credit Card Default Dataset," J. Adv. Comput. Syst., vol. 4, no. 6, pp. 74-85, Jun. 2024, doi: 10.69987/JACS.2024.40606.

[43] Y. Chen, Y. Zhang, D. Chau, and M. Sherman, "Credit Card Default Risk Tiering with Probability Calibration and Uncertainty-Driven Rejection: A Reproducible Study on the UCI Credit Card Clients Dataset," J. Adv. Comput. Syst., vol. 3, no. 4, pp. 31-47, Apr. 2023, doi: 10.69987/JACS.2023.30403.

[44] Z. Li, S. Zhou, and Z. Zhou, "Financial Risk Dashboard Design for Institutional RWA Investors: Visual Hierarchy, Chart Comprehension, and Explainability in FinChart-Bench," Int. J. Graph. Des., vol. 3, no. 1, pp. 196-210, May 2025, doi: 10.51903/ijgd.v3i1.3715.

[45] Q. Xin, "Uncertainty-Aware Late Fusion for 3D Perception (Confidence Calibration + Fusion Rule Learning)," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 215-238, Apr. 2025, doi: 10.51903/jtie.v4i1.485.

[46] S. He, H. Tu, and I. Liu, "Safe PD Capacity Forecasting with Time-Series Foundation Models and Calibrated Uncertainty for Heterogeneous GPU Clusters," J. Adv. Comput. Syst., vol. 3, no. 4, pp. 48-66, Apr. 2023, doi: 10.69987/JACS.2023.30404.

[47] J. Jin, "Calibrated Resume-Job Matching for Trustworthy LLM-Assisted Recruiter Screening: Pairwise Matching, Probability Calibration, and Selective Refusal on Two Public Recruitment Datasets," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 625-648, Dec. 2025, doi: 10.51903/jtie.v4i3.529.

[48] S. Chen, S. He, and E. Sun, "Risk-Bounded GPU Resource Oversubscription via Conformal Demand Envelopes in Production AI Clusters," J. Adv. Comput. Syst., vol. 4, no. 5, pp. 119-134, May 2024, doi: 10.69987/JACS.2024.40509.

[49] Z. S. Zhong, C. Li, and H. Rao, "Trajectory Reliability Prediction for Generalist AI Agents: Tool-Use Failure Analysis and Success Forecasting on ZClawBench," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 341-360, Apr. 2026, doi: 10.51903/jtie.v5i1.539.

[50] S. He, X. Chang, and E. Sun, "Cross-Cloud Transfer Learning for AI Training Capacity Forecasting under Workload and Topology Distribution Shift," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 100-120, Jan. 2024, doi: 10.69987/JACS.2024.40108.

[51] Q. Xin, "Probabilistic Bike-Sharing Demand Forecasting under Changing Weather and Seasonal Regimes with Transformer-Based Models," Transport Findings, Mar. 2026, doi: 10.32866/001c.157499.

[52] J. Jin, "LLM-Style Evidence Cards for Scientific Search Interfaces: A UI/UX Design Framework for Retrieval Transparency, Ranking Trust, and Visual Evidence Hierarchy," Int. J. Graph. Des., vol. 3, no. 2, pp. 397-414, Oct. 2025, doi: 10.51903/ijgd.v3i2.3698.

[53] G. Liu, S. He, and H. Wong, "LLM-Compatible Visual Brief Cards for AI Infrastructure Capacity Dashboards: A UI/UX Framework for Turning Forecast Risk into Graphic Design Decisions," Int. J. Graph. Des., vol. 3, no. 1, pp. 196-213, May 2025, doi: 10.51903/ijgd.v3i1.3723.

[54] W. Su, H. Rao, and E. Ma, "Privacy and Data-Integrity Risk Cards for LLM Agents: A UI/UX Design Framework for Secure Human Oversight under Prompt-Injection Attacks," Int. J. Graph. Des., vol. 4, no. 1, pp. 186-191, Apr. 2026, doi: 10.51903/ijgd.v4i1.3699.

[55] Y. Li, S. Lu, and L. Zhao, "LLM-as-Design-Critic: Aligning AI-Generated UI Feedback with Human Graphic Design Judgment," Int. J. Graph. Des., vol. 3, no. 1, pp. 196-215, May 2025, doi: 10.51903/ijgd.v3i1.3661.

[56] J. Nie, G. Liu, C. Li, and T. Zou, "Evidence-Constrained Incident Visualization Cards for Distributed Cloud Logs: A UI/UX Framework for Turning Hadoop, OpenStack, and ZooKeeper Logs into Actionable SRE Design Interfaces," Int. J. Graph. Des., vol. 4, no. 1, pp. 179-185, Apr. 2026, doi: 10.51903/ijgd.v4i1.3703.

[57] Q. Xin, "Hybrid Cloud Architecture for Efficient and Cost-Effective Large Language Model Deployment," J. Inf. Syst. Informatics, vol. 7, no. 3, pp. 2182-2195, Sep. 2025, doi: 10.51519/journalisi.v7i3.1170.

[58] D. Zheng, B. Zhang, and J. Geibel, "VerifySafe: Toxicity-Safe Agent Responses under Adversarial Prompts with Evidence-Based Self-Verification," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 67-82, Jan. 2024, doi: 10.69987/JACS.2024.40106.

[59] Q. Xin, "Explaining OpenStack Failure-Injection Log Anomalies with Retrieved Normal Prototypes," Emerg. Inf. Sci. Technol., vol. 6, no. 2, Nov. 2025, doi: 10.18196/eist.v6i2.31232.

[60] D. Zheng and C. Li, "Behavior-Level Jailbreak Resistance via Multi-Stage Refusal + Utility Preservation," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 83-99, Jan. 2024, doi: 10.69987/JACS.2024.40107.

[61] Q. Xin, "Log Anomaly Detection with Conformal Alert Control and Evidence-Grounded Incident Ticket Generation," AVITEC, vol. 8, no. 2, pp. 247-264, May 2026, doi: 10.28989/avitec.v8i2.3974.

[62] D. Zheng, C. Li, and H. Davidson, "Continual Red-Teaming for In-the-Wild Jailbreaks via Online Guardrail Updates and Guardrail Distillation," J. Adv. Comput. Syst., vol. 3, no. 2, pp. 35-49, Feb. 2023, doi: 10.69987/JACS.2023.30203.

[63] Q. Xin, "LiDAR-Camera Object-Level Fusion for Multi-Target Tracking Using JPDA and EKF: A

Reproducible Empirical Study on a PandaSet-Parameterised Five-Sequence Dataset," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 54-76, Apr. 2026, doi: 10.51903/jtie.v5i1.486.

[64] Y. Li, "Findable then Explainable: Retrieval-Summary Integration for Code Intelligence on a Lightweight CodeSearchNet Subset," J. Adv. Comput. Syst., vol. 4, no. 7, pp. 65-82, Jul. 2024, doi: 10.69987/JACS.2024.40706.

[65] [65] Q. Xin, "Host-Based Intrusion Detection with System Call Sequences: Window Localization and Forensic Narratives," AVITEC, vol. 8, no. 2, pp. 325-334, Jun. 2026, doi: 10.28989/avitec.v8i2.3973.