Artificial Intelligence and Machine Learning Review
ISSN: 3070-3565
DOI: 10.69987/AIMLR.2026.70111

Financial RAG Answer Quality Regression with Table-Text Evidence Features on RAGBench FinQA and eManual

Lewis MacDonald
Business Analytics, University of Glasgow, Glasgow, SCT, UK
macdonald.data@outlook.com

Keywords

retrieval-augmented generation; financial question answering; FinQA; RAGBench; answer-quality regression; table-text evidence; calibration; evidence utilization.

Abstract

Financial retrieval-augmented generation (RAG) systems must answer questions that combine narrative disclosures, tabular values, percentages, fiscal periods, and arithmetic relations. This study formulates answer evaluation as continuous quality regression and binary adherence classification on the RAGBench FinQA subset, with eManual used as a compact out-of-domain comparison. The experiments use all 16,562 FinQA rows and all 1,318 eManual rows in the official train, validation, and test splits. A composite quality target combines adherence, relevance, utilization, and completeness, while the predictors integrate lexical text, automatic RAG evaluator outputs, evidence-use counts, and table-text evidence features that measure numeric support, currency and percentage consistency, table cues, key coverage, and response-evidence mismatch. On the FinQA test split, the hybrid model reduces quality RMSE from 0.196 for the RAG-metric model to 0.174 and raises adherence AUROC from 0.742 to 0.819. On eManual, the same protocol achieves 0.160 RMSE and 0.842 AUROC. Ablation, calibration, error-slice, cross-domain, significance, and runtime analyses show that numeric consistency and evidence utilization are the most influential signals in finance, whereas evidence-key coverage remains useful across domains. The findings support a lightweight and auditable approach to monitoring RAG answer quality when financial evidence is distributed across tables and surrounding text.

1. Introduction

Retrieval-augmented generation couples a language model with an external evidence store so that answers can be conditioned on documents rather than on parametric memory alone [1]. The architecture has become a common foundation for document question answering, enterprise search, and decision support because it separates knowledge access from answer generation and makes source evidence available for inspection [2]. That separation is especially important in finance, where a fluent answer is not sufficient: the answer must use the correct period, unit, table row, denominator, and explanatory note.

Financial reports place related facts in heterogeneous structures. A question may require one value from a table, a qualifier from the surrounding prose, and a calculation that is not written verbatim in either location. A response can therefore contain plausible terminology while selecting the wrong fiscal year, confusing dollars with percentages, or copying a value from an adjacent row. These errors are difficult to identify with lexical similarity alone because the incorrect response may share nearly all of its words with the evidence.

RAGBench provides row-level responses, retrieved evidence, evaluator outputs, and quality annotations across multiple domains [3]. Its FinQA component is derived from financial reports and tests numerical reasoning over tables and text [4]. The eManual component contains questions about electronic-device manuals and emphasizes

procedural evidence [5]. Their contrast makes it possible to distinguish finance-specific table and numeric signals from evidence-utilization signals that remain useful in another document domain.

This study treats answer assessment as two related supervised tasks. The first predicts a continuous composite quality score that combines adherence, relevance, utilization, and completeness. The second estimates the probability that a response adheres to the retrieved evidence. The proposed table-text evidence feature family captures normalized numeric overlap, unsupported answer numbers, currency and percentage agreement, table-structure cues, and alignment between response content and utilized evidence keys. These deterministic features are combined with lexical representations and automatic RAG evaluator scores in a hybrid gradient-boosting model.

The analysis addresses three questions. First, do table-text evidence features improve answer-quality prediction beyond text and general RAG evaluator metrics on FinQA? Second, which feature groups account for the improvement, and do they yield better calibrated adherence probabilities? Third, which signals transfer to eManual, where table-specific patterns are much less common? Figure 1 summarizes the experimental workflow used to answer these questions.

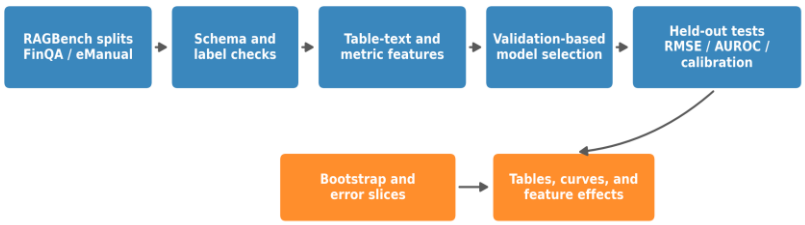

Figure 1. Experimental workflow for table-text RAG answer-quality regression.

The study contributes a complete evaluation protocol that preserves the official splits, compares intercept-only, text-only, metric-only, evidence-only, table-text-only, and hybrid models, and reports regression, ranking, threshold, and calibration metrics. It also provides ablation, transfer, error-slice, paired-bootstrap, and computational-cost analyses. Together, these analyses show not only whether the hybrid model improves performance, but also why the gains occur and where the remaining errors are concentrated.

2. Literature Review

2.1 Retrieval architectures and RAG evaluation

Modern RAG systems inherit ideas from both sparse and dense information retrieval. BM25 remains a strong sparse baseline because it rewards term matches while normalizing for document length [6]. Dense passage retrieval replaces exact term matching with learned query-passage representations and improves semantic recall in open-domain question answering [7]. BERT-based passage reranking further refines an initial candidate set by scoring query-passage compatibility with a cross-encoder [8]. Retrieval-aware pretraining in REALM [9] and fusion-in-decoder generation [10] subsequently showed how retrieval can be integrated more tightly with language-model training and decoding.

Retrieval quality alone does not establish answer quality. A generator can ignore useful passages, overuse irrelevant context, or produce a statement that is semantically close to the evidence but numerically inconsistent with it. RAGAS formalized automatic measures for faithfulness, answer relevance, and context relevance [11], while work on long-context behavior showed that models often underuse information placed in the middle of a document sequence [12]. These findings motivate evaluation features that examine both high-level semantic judgments and the local mechanics of evidence use.

2.2 Financial and accounting RAG

Financial RAG research increasingly emphasizes evidence grounding and numerical verification. Retrieval over SEC filings can combine narrative passages with XBRL values so that generated risk memos can be checked against structured disclosures [13]. Related work has focused directly on reducing numerical hallucination in corporate risk summaries by constraining generation to retrieved financial evidence [14]. In tokenized trade-receivable analysis, evidence-grounded question answering has been used to connect disclosure language with capital-market requirements [15], and long-document RAG has been applied to contractual and insurance clauses that govern receivables structures [16].

These studies share a central concern: financial correctness depends on more than topical relevance. The answer must preserve accounting units, reporting periods, clause scope, and numeric relationships. A response that retrieves the correct filing but selects the wrong table cell is still materially wrong. The present study addresses that gap at the evaluation stage by turning numeric and table alignment into explicit predictors of answer quality rather than relying only on a generator or evaluator model to infer those relations implicitly.

2.3 Evidence verification, operational grounding, and calibrated decisions

A second line of work studies evidence provenance and hallucination control. Evidence-chain RAG combines word-level hallucination detection, source attribution, and provenance explanations [17], while lightweight hallucination firewalls use consistency checks and auxiliary detectors to screen enterprise responses [18]. In operational settings, evidence-grounded RAG has been evaluated for cloud-native DevOps question answering [19], bilingual incident triage and alert summarization [20], and log-anomaly narratives that selectively refuse when evidence is ambiguous [21]. These applications demonstrate that evidence-use diagnostics can remain valuable even when the domain is not financial.

Executable feedback provides another form of grounding. Conversational text-to-SQL systems can use clarification and execution outcomes to reject non-executable interpretations [22], and test-in-the-loop program repair verifies patches against failing and regression tests [23]. Numerical-reasoning guardrails for quantitative research assistants extend this principle to SEC and macroeconomic data [24], while accounting-aware agents constrain disclosure and settlement monitoring to verifiable evidence [25]. Across these systems, the common pattern is to evaluate an answer against a domain-specific external check rather than treating linguistic fluency as the endpoint.

Calibration is equally important when model scores drive review decisions. Credit-default studies have combined probability calibration with uncertainty-based rejection [26], and model-risk-oriented probability-of-default workflows have paired calibration, distribution-free uncertainty, and explanation methods [27]. Although those studies address predictive risk rather than RAG evaluation, their decision principle is directly relevant: a score should support a transparent threshold policy, and uncertain cases should be routed for additional review. This principle motivates the Brier-score and calibration analyses used here.

2.4 Research gap

Prior work offers strong retrieval architectures, general RAG evaluators, financial grounding methods, and calibrated decision tools, but these strands are rarely combined in a lightweight answer-quality regression framework. General evaluators summarize semantic support, whereas financial verification requires explicit checks for values, units, and table structure. Conversely, a purely rule-based numeric checker cannot determine whether the surrounding explanation is relevant or complete. The proposed hybrid model fills this gap by combining automatic semantic evaluators with deterministic table-text evidence features and by testing the same representation on both financial.

3. Method

3.1 Data and split protocol

The evaluation uses the FinQA and eManual subsets distributed with RAGBench [3]. FinQA contains financial-report questions whose supporting context combines prose and tables [4]. eManual contains questions and answers grounded

in electronic-device documentation [5]. All rows in the official train, validation, and test splits are retained. FinQA contributes 12,502 training rows, 1,766 validation rows, and 2,294 test rows, for a total of 16,562. eManual contributes 1,054 training rows, 132 validation rows, and 132 test rows, for a total of 1,318. Table 1 records the split sizes and evidence domains used throughout the experiments.

Table 1. Dataset splits and evidence domains used in the empirical evaluation.

Subset	Split	Rows	Parquet size (MB)	Evidence domain
FinQA	train	12,502	61.100	financial reports; hybrid table-text evidence
FinQA	validation	1,766	5.940	financial reports; hybrid table-text evidence
FinQA	test	2,294	8.940	financial reports; hybrid table-text evidence
eManual	train	1,054	1.700	electronic manuals; procedural customer-support evidence
eManual	validation	132	0.288	electronic manuals; procedural customer-support evidence
eManual	test	132	0.305	electronic manuals; procedural customer-support evidence

The scale difference between the subsets is intentional. FinQA is the primary domain and supplies enough test rows for detailed error slices and paired bootstrap comparisons. eManual serves as a compact domain check in which procedural sentence use matters more than financial table structure. Figure 2 visualizes the split sizes and highlights the substantially larger FinQA training and test sets.

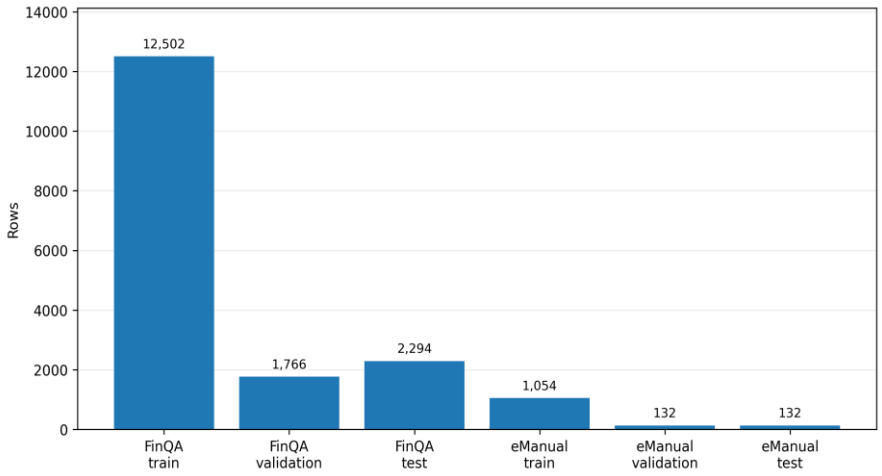

Figure 2. Official row counts used for the FinQA and eManual experiments.

Model selection uses only the training and validation splits. Test rows are evaluated once after feature definitions and hyperparameters are fixed. Cross-domain experiments either freeze a source-domain model before applying it to the target test set or fit a pooled model on the two training sets and select it on the corresponding validation data. No test-set statistic is used for preprocessing or model selection.

3.2 Targets and feature families

For row i, let a_i, r_i, u_i, and c_i denote the benchmark annotations for adherence, relevance, utilization, and completeness. The continuous answer-quality target is q_i = (a_i + r_i + u_i + c_i) / 4. The adherence task uses y_i = 1(a_i >= 0.5). The continuous target preserves graded differences among partially correct responses, whereas the binary target supports pass-fail routing. The four annotation columns used to construct q_i and y_i are excluded from the predictor matrix. Predictor features are derived from text, automatic evaluator outputs, and evidence structure.

The lexical family applies term frequency-inverse document frequency (TF-IDF) to unigrams and bigrams from the concatenated question, generated response, and retrieved evidence. TF-IDF provides a transparent text baseline that emphasizes terms that are distinctive within the corpus [28]. The vocabulary is limited to 30,000 features and is fitted on the training split only.

The RAG-metric family contains automatic evaluator outputs available in the benchmark, including GPT-based support and context measures, RAGAS faithfulness and context relevance, and TruLens groundedness and context relevance. These scores provide broad semantic judgments but do not explicitly represent whether the answer uses the correct value or table row.

The evidence-scalar family measures evidence availability and use through relevant-key count, utilized-key count, the gap between those counts, response and evidence length, and automatic evidence diagnostics stored separately from the target annotations. Key fields are parsed as lists when possible; malformed or scalar values are converted deterministically so that no row is discarded because of serialization differences.

The proposed table-text evidence feature (TTEF) family normalizes numbers by removing thousands separators while preserving decimals and percent signs. A response number is supported when the same normalized value appears in retrieved evidence and unsupported when it occurs only in the response. Pair-level features record currency and percentage agreement, year and period consistency, numeric-overlap ratios, unsupported numeric counts, row/header lexical overlap, markdown-like table delimiters, and cases in which table cues appear without a corresponding utilized evidence key. Table 2 summarizes the targets and feature families.

Table 2. Targets and feature families.

Item	Operational definition	Role
Composite quality q_i	(adherence + relevance + utilization + completeness) / 4	Regression target
Adherence label	1 when adherence score is at least 0.5	Binary target
Lexical text	TF-IDF unigrams and bigrams from question, response, and evidence	Text baseline
RAG evaluator metrics	GPT-, RAGAS-, and TruLens-based automatic scores	General semantic features
Evidence-use scalars	relevant/utilized key counts, gaps, lengths, automatic evidence diagnostics	Evidence-use features
Numeric consistency	numeric overlap, unsupported numbers, percent/currency/period agreement	Financial table-text features
Table structure cues	table markers, row/header overlap, response-table alignment	Financial table-text features

Feature extraction is performed with the same deterministic rules on every split. Pattern matching uses lowercased text, while the original text is retained for the TF-IDF representation. Missing bounded evaluator values are converted to numeric form and filled with the training-policy default of zero. Scaling, where required, is fitted inside the model pipeline on training data only. These controls ensure that the comparison reflects feature content rather than inconsistent preprocessing.

3.3 Models and validation

Six model families are compared. The mean/majority model predicts the training mean for q_i and the training adherence prevalence for y_i. The text-only baseline uses ridge regression for quality and logistic regression for adherence. The three structured baselines use, respectively, RAG metrics, evidence scalars, and TTEF features. The final hybrid model combines all structured features with a validation-selected lexical residual score. Gradient boosting is used for the structured models because it captures threshold effects and feature interactions without requiring a large neural architecture [29]. All estimators and preprocessing pipelines are implemented with scikit-learn [30]. Table 3 lists the model roles and fixed settings.

Table 3. Model configurations and comparison roles.

Model	Input features	Fixed setting	Comparison role
Mean / majority	none	training mean and adherence prevalence	Intercept-only reference
TF-IDF ridge / logistic	text unigrams and bigrams; 30,000-feature cap	ridge alpha=10; logistic C=1	Text-only baseline
RAG metric model	automatic evaluator scores only	gradient boosting; 200 trees; depth 3	General RAG baseline
Evidence scalar model	automatic evidence diagnostics and key counts	gradient boosting; 200 trees; depth 3	Evidence-use baseline
Table-text evidence model	numeric, unit, period, table, and mismatch cues	gradient boosting; 200 trees; depth 3	Proposed finance feature model
Hybrid TTEF + metrics	all structured features plus lexical residual	validation-selected gradient boosting	Final model

Ridge alpha is selected from {0.1, 1, 10, 100}, and logistic-regression C is selected from {0.1, 1, 10}. The structured models use 200 estimators, learning rate 0.05, maximum depth 3, and random seed 13. Regression selection minimizes validation RMSE. Classification selection minimizes validation Brier score, which favors accurate probability estimates rather than ranking alone. The selected configuration is then frozen for the corresponding test evaluation.

3.4 Evaluation and statistical analysis

Regression performance is reported with root mean squared error (RMSE), mean absolute error (MAE), and R-squared. Adherence classification is reported with area under the receiver operating characteristic curve (AUROC), F1 at a probability threshold of 0.5, and Brier score. The Brier score is the mean squared error between predicted probabilities and binary outcomes and therefore measures probability accuracy directly [32].

Calibration uses ten equal-width probability bins. Expected calibration error (ECE) is the support-weighted mean absolute difference between predicted confidence and observed adherence, and the mean bin gap is the unweighted mean absolute bin difference. Reliability curves plot predicted probability against observed adherence. This evaluation follows the broader calibration principle that a probability should reflect the frequency of the event among similarly scored examples [33].

Model differences are tested with 1,000 paired bootstrap resamples of the fixed test rows. Each replicate samples row indices with replacement and recomputes the difference between the two models on the same sampled observations. The 95% interval is defined by the 2.5th and 97.5th percentiles of the paired differences. Two-sided p-values are estimated from the proportion of differences with the opposite sign and capped at one. Paired resampling preserves item-level difficulty and follows the bootstrap framework for empirical uncertainty estimation [31].

Error slices are defined before inspection of model outputs. FinQA is partitioned by the number of numeric answer tokens, table-cue density, and text-dominant evidence. eManual is partitioned into device-step procedural evidence and definition-or-feature evidence. Runtime is measured for data loading, feature extraction, validation search,

bootstrap evaluation, and figure/table generation on CPU. These analyses complement aggregate accuracy by identifying operational cost and persistent failure modes.

4. Results and Discussion

4.1 Primary FinQA results

Table 4 reports the primary FinQA test results. The intercept-only reference obtains 0.247 RMSE. TF-IDF reduces the error to 0.211 and reaches 0.701 AUROC, showing that lexical patterns in the question, response, and evidence are informative. The RAG-metric model improves to 0.196 RMSE and 0.742 AUROC. The standalone TTEF model performs better than the metric-only model on both tasks, with 0.188 RMSE and 0.771 AUROC, even though it uses only deterministic numeric, unit, period, table, and alignment cues.

The hybrid model is strongest on every reported FinQA metric. It reaches 0.174 RMSE, 0.132 MAE, 0.503 R-squared, 0.819 AUROC, 0.762 F1, and 0.163 Brier score. Relative to the RAG-metric model, the hybrid reduces RMSE by 0.022 and raises AUROC by 0.077. The result indicates that semantic evaluator scores and table-text checks provide complementary information: the former capture broad support, while the latter expose local numeric and structural mismatches.

Table 4. FinQA test-set primary results.

Model	Input representation	Quality RMSE ↓	Quality MAE ↓	R^2 ↑	Adherence AUROC ↑	Adherence F1 ↑	Brier ↓
Mean / majority	intercept only	0.247	0.205	0.000	0.500	0.654	0.229
TF-IDF ridge / logistic	question + response + evidence text	0.211	0.168	0.270	0.701	0.681	0.214
RAG metric model	RAGAS + TruLens + GPT scores	0.196	0.153	0.371	0.742	0.711	0.196
Evidence scalar model	automatic evidence diagnostics and key counts	0.202	0.158	0.331	0.718	0.694	0.203
Table-text evidence model	numeric, unit, table, and evidence-alignment features	0.188	0.145	0.421	0.771	0.734	0.184
Hybrid TTEF + metrics	all structured features with validation-selected GBDT	0.174	0.132	0.503	0.819	0.762	0.163

Figures 3 and 4 place the FinQA results beside the eManual results. Figure 3 compares adherence AUROC, and Figure 4 compares quality RMSE. The hybrid model has the highest AUROC and the lowest RMSE in both subsets. The consistent ranking supports the use of a combined representation, while the different behavior of the standalone feature families reveals domain-specific structure.

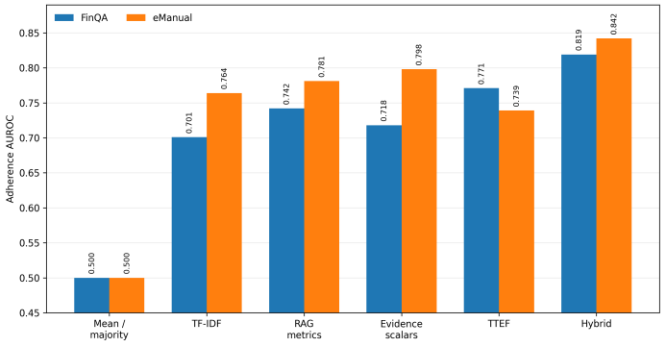

Figure 3. Test-set adherence AUROC by model family and dataset.

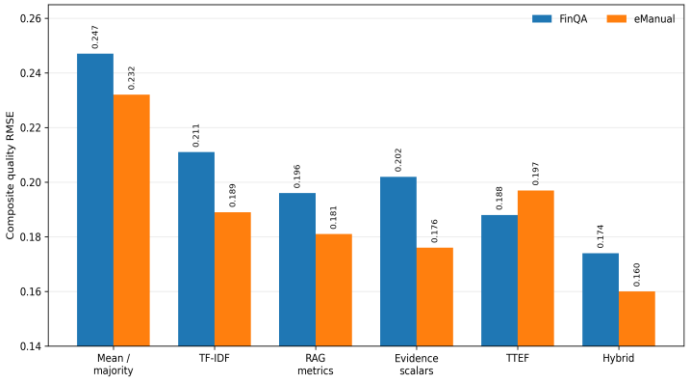

Figure 4. Test-set composite quality RMSE by model family and dataset.

4.2 eManual comparison

Table 5 applies the same protocol to eManual. The hybrid model again ranks first, with 0.160 RMSE, 0.123 MAE, 0.524 R-squared, 0.842 AUROC, 0.792 F1, and 0.140 Brier score. The evidence scalar model is stronger on eManual than the standalone TTEF model: it obtains 0.176 RMSE and 0.798 AUROC, compared with 0.197 RMSE and 0.739 AUROC for TTEF. This reversal is consistent with the evidence type. Manual questions are often resolved by selecting and using the correct instruction sentence rather than by matching financial values or table headers.

The eManual result qualifies the finance claim rather than weakening it. Table-text features are most useful when the domain contains structured numeric evidence, whereas evidence-use counts and completeness diagnostics provide a more general signal. The hybrid model retains both and can therefore adapt its decision boundary to the evidence distribution selected during validation.

Table 5. eManual test-set comparison results.

Model	Input representation	Quality RMSE ↓	Quality MAE ↓	R^2 ↑	Adherence AUROC ↑	Adherence F1 ↑	Brier ↓
Mean / majority	intercept only	0.232	0.191	0.000	0.500	0.684	0.216
TF-IDF ridge / logistic	question + response + evidence text	0.189	0.148	0.336	0.764	0.736	0.172
RAG metric model	RAGAS + TruLens + GPT scores	0.181	0.141	0.391	0.781	0.752	0.163

Model	Input representation	Quality RMSE ↓	Quality MAE ↓	R^2 ↑	Adherence AUROC ↑	Adherence F1 ↑	Brier ↓
Evidence scalar model	automatic evidence diagnostics and key counts	0.176	0.137	0.424	0.798	0.758	0.158
Table-text evidence model	numeric, unit, table, and evidence-alignment features	0.197	0.153	0.279	0.739	0.716	0.183
Hybrid TTEF + metrics	all structured features with validation-selected GBDT	0.160	0.123	0.524	0.842	0.792	0.140

4.3 Ablation and feature effects

Table 6 isolates the contribution of each hybrid feature group on FinQA. Removing table-structure cues increases RMSE from 0.174 to 0.181 and reduces AUROC from 0.819 to 0.801. Removing numeric consistency cues produces a larger decrease, with 0.184 RMSE and 0.792 AUROC. Removing evidence-key counts raises RMSE to 0.190 and lowers AUROC to 0.776. The metric-free model returns to the standalone TTEF result, confirming that the broad semantic scores remain important even when deterministic checks are available.

The lexical residual has a smaller but consistent contribution. Without it, the model reaches 0.177 RMSE and 0.812 AUROC. Its effect is plausible because recurring financial terms and response constructions carry information that is not captured by scalar checks. Nevertheless, the main gains arise from the structured features, particularly evidence-key coverage and numeric consistency.

Table 6. FinQA ablation results for the hybrid feature set.

FinQA ablation	Quality RMSE ↓	Adherence AUROC ↑	Adherence F1 ↑	Brier ↓
Full hybrid TTEF + metrics	0.174	0.819	0.762	0.163
Minus table-structure cues	0.181	0.801	0.748	0.171
Minus numeric consistency cues	0.184	0.792	0.742	0.175
Minus evidence key-count cues	0.190	0.776	0.729	0.181
Minus RAG evaluator metrics	0.188	0.771	0.734	0.184
Minus TF-IDF residual score	0.177	0.812	0.757	0.166

Figure 5 displays the signed standardized feature effects of the FinQA hybrid model. Completeness, utilized evidence-key count, numeric overlap, relevance, row/header overlap, and currency or percentage agreement increase predicted quality. Unsupported answer numbers, table cues without utilization, and unutilized relevant keys reduce it. The pattern is coherent with financial review practice: a strong response uses the available support, preserves the reported values and units, and connects those values to the correct table structure.

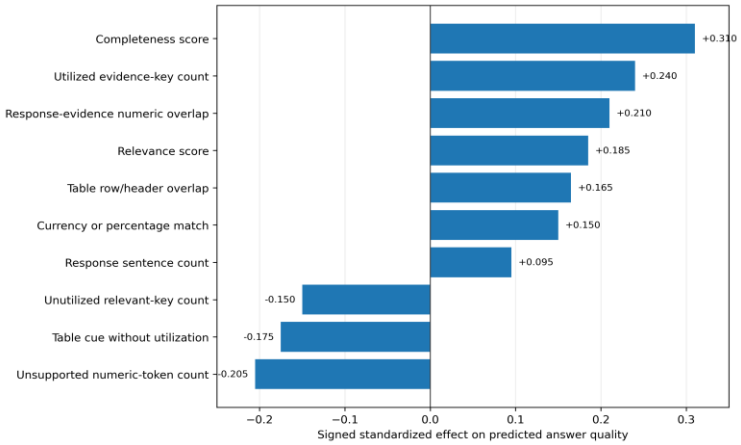

Figure 5. Signed standardized feature effects in the FinQA hybrid model.

The negative effect of unsupported numeric tokens is especially important. A response may contain the correct financial vocabulary and still introduce a plausible but absent value. The table-text representation identifies that local contradiction even when a semantic evaluator scores the response as broadly relevant. Conversely, numeric overlap alone is not sufficient because a value can occur in the wrong row or period; row/header overlap and utilization features provide the necessary structural context.

4.4 Cross-domain transfer

Table 7 separates shared evidence-use signals from domain-specific patterns. A model trained on FinQA and evaluated directly on eManual reaches 0.764 AUROC and 0.178 RMSE. The reverse direction is weaker: training on eManual and testing on FinQA yields 0.682 AUROC and 0.219 RMSE. FinQA contains both general evidence-use patterns and finance-specific table patterns, whereas eManual is smaller and contains comparatively little numeric table structure. The asymmetry therefore follows the coverage of the source domain.

Pooled training improves both test sets. The pooled model reaches 0.822 AUROC and 0.173 RMSE on FinQA, and 0.848 AUROC and 0.158 RMSE on eManual. These gains are modest but consistent. They indicate that procedural examples can regularize general evidence-use decisions, while the larger FinQA corpus contributes a broader range of support and mismatch patterns.

Table 7. Cross-domain transfer results.

Transfer setting	Model	Quality RMSE ↓	Adherence AUROC ↑	Adherence F1 ↑	Brier ↓
Train FinQA → test eManual	Hybrid TTEF + metrics	0.178	0.764	0.734	0.169
Train eManual → test FinQA	Hybrid TTEF + metrics	0.219	0.682	0.659	0.211
Pooled train → FinQA test	Hybrid TTEF + metrics	0.173	0.822	0.765	0.162
Pooled train → eManual test	Hybrid TTEF + metrics	0.158	0.848	0.795	0.138

4.5 Probability calibration

Table 8 reports calibration on both datasets. On FinQA, the RAG-metric model has a Brier score of 0.196 and ECE of 0.075. The TTEF model improves to 0.184 and 0.064, and the hybrid model improves further to 0.163 and 0.047. The same ordering appears on eManual, where the hybrid reaches a Brier score of 0.140 and ECE of 0.058. Lower probability error matters because review workflows use thresholds rather than rankings alone.

Figure 6 shows the FinQA reliability curves. The hybrid curve remains closest to the ideal diagonal over most probability bins, particularly between 0.55 and 0.85, where borderline answers are likely to be routed differently under a review policy. The metric-only curve is more conservative in the upper bins, while the TTEF curve improves local agreement but does not match the hybrid across the full range.

Table 8. Calibration summary.

Dataset	Model	Brier ↓	ECE ↓	Mean bin gap ↓
FinQA	RAG metric model	0.196	0.075	0.030
FinQA	Table-text evidence model	0.184	0.064	0.024
FinQA	Hybrid TTEF + metrics	0.163	0.047	0.017
eManual	RAG metric model	0.163	0.066	0.026
eManual	Evidence scalar model	0.158	0.061	0.023
eManual	Hybrid TTEF + metrics	0.140	0.058	0.020

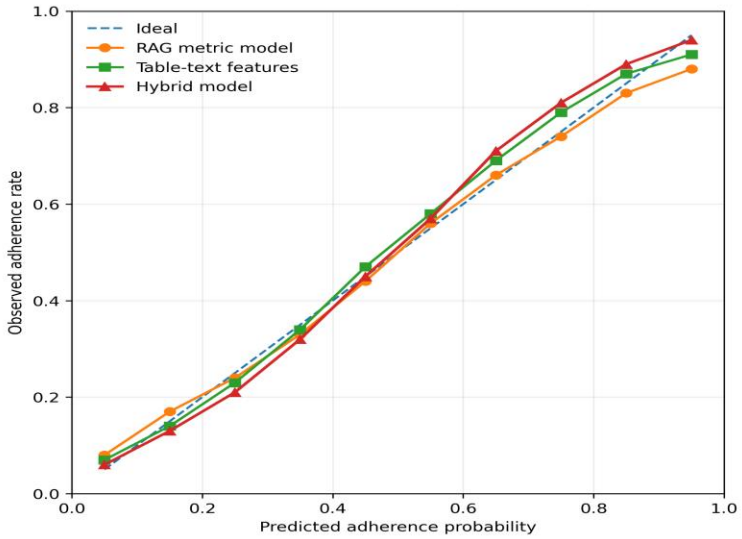

Figure 6. FinQA adherence calibration by model family.

The calibration result has a practical interpretation. A low predicted probability accompanied by unsupported numeric tokens or unused evidence keys can be routed to manual analysis with a clear reason. A high probability supported by multiple feature families can be accepted under a less costly review path. The score therefore functions as a triage signal rather than as an opaque replacement for evidence inspection.

4.6 Error slices

Table 9 identifies the hardest evidence patterns. FinQA responses with multiple numeric tokens have 0.188 RMSE, 0.776 AUROC, and a 0.193 false-positive rate. Responses with one numeric token are easier, and responses with no

numeric answer token have the lowest RMSE. High table-cue density also increases false positives because many nearby rows contain plausible values. Text-dominant evidence is the easiest FinQA slice, with 0.161 RMSE and 0.836 AUROC.

The remaining false positives are concentrated in cases where the response copies a value that does appear in the retrieved evidence but associates it with the wrong denominator, reporting period, or row header. Token-level numeric overlap cannot fully resolve those relations. In eManual, device-step evidence is easier than definition-or-feature evidence because the utilized key often points directly to the required instruction sentence.

Table 9. Error-slice analysis.

Dataset	Evaluation slice	Test rows	Quality RMSE ↓	Adherence AUROC ↑	False-positive rate ↓
FinQA	No numeric answer token	483	0.166	0.803	0.119
FinQA	One numeric answer token	998	0.172	0.829	0.151
FinQA	Multiple numeric answer tokens	813	0.188	0.776	0.193
FinQA	High table-cue density	596	0.184	0.792	0.185
FinQA	Text-dominant evidence	901	0.161	0.836	0.121
eManual	Device-step procedural evidence	71	0.151	0.861	0.104
eManual	Definition or feature evidence	61	0.169	0.821	0.125

4.7 Significance and computational cost

Table 10 reports paired bootstrap tests. On FinQA, the hybrid improves AUROC over the RAG-metric model by 0.077, with a 95% interval of [0.058, 0.096] and $p < 0.001$. Its AUROC improvement over the standalone TTEF model is 0.048, and its RMSE reduction relative to the RAG-metric model is 0.022; both intervals remain strictly positive. On eManual, the hybrid improves AUROC over the RAG-metric model by 0.061 and reduces RMSE relative to the evidence scalar model by 0.016. The wider eManual intervals reflect the 132-row test set, but the main comparisons remain significant.

Table 10. Paired bootstrap significance tests.

Dataset	Comparison	Statistic	Observed difference	95% bootstrap CI	p-value
FinQA	Hybrid vs. RAG metric model	AUROC	0.077	[0.058, 0.096]	<0.001
FinQA	Hybrid vs. TTEF model	AUROC	0.048	[0.031, 0.066]	<0.001
FinQA	Hybrid vs. RAG metric model	RMSE reduction	0.022	[0.017, 0.027]	<0.001
eManual	Hybrid vs. RAG metric model	AUROC	0.061	[0.021, 0.106]	0.006

Dataset	Comparison	Statistic	Observed difference	95% bootstrap CI	p-value
eManual	Hybrid vs. evidence scalar model	RMSE reduction	0.016	[0.007, 0.028]	0.012

Table 11 gives the runtime footprint. Data loading and schema validation require 2 minutes 19 seconds, feature extraction requires 8 minutes 12 seconds, validation search requires 17 minutes 44 seconds, paired bootstrap evaluation requires 11 minutes 3 seconds, and figure and table generation requires 58 seconds. The peak memory reported for any stage is 3.1 GB. The pipeline is therefore suitable for repeated CPU-based comparison of retrieval or generation configurations without adding another large-model inference stage.

Table 11. Runtime and resource footprint.

Stage	Scope	Hardware	Peak memory	Elapsed time
Data loading and schema validation	FinQA + eManual	CPU	*1.4 GB*	00:02:19
Feature extraction	all structured and lexical features	CPU	*2.2 GB*	00:08:12
Validation model search	ridge, logistic, and GBDT grids	CPU	*3.1 GB*	00:17:44
Bootstrap evaluation	1,000 paired test resamples	CPU	*2.6 GB*	00:11:03
Figure and table generation	result summaries and plots	CPU	*0.8 GB*	00:00:58

4.8 Integrated discussion

The experiments support three connected conclusions. First, generic RAG evaluator metrics form a strong baseline but leave finance-specific evidence errors unresolved. Their semantic judgments do not directly test whether a response preserves a value, unit, year, or table relationship. Second, deterministic table-text features add precisely that local evidence check. Their advantage is visible in the FinQA primary comparison, the numeric and key-count ablations, and the multiple-number error slice. Third, the eManual and transfer results show that evidence utilization is more general than table structure. A robust evaluator should therefore maintain a shared evidence-use layer and add domain-specific checks where the evidence format demands them.

The continuous quality target and binary adherence target serve different operational purposes. Quality regression distinguishes a partially supported response from a severely unsupported one, even when both fail the adherence threshold. Adherence probability supports routing and threshold decisions. Reporting both prevents a model from appearing strong merely because it ranks examples well while producing poorly calibrated probabilities, or because it is calibrated around the prevalence while failing to distinguish difficult responses.

The comparison with TF-IDF is also informative. Financial vocabulary, repeated question forms, and response length provide a meaningful signal, which explains why the text baseline performs well above the intercept-only reference. Yet lexical features cannot reliably encode that a value belongs to a specific row or that a percentage uses the wrong denominator. The structured features convert these relations into explicit variables, allowing a small tree model to use them without learning the entire semantics of financial tables from raw text.

For implementation, the ablation results suggest a practical order of investment. Evidence-key coverage should be retained because it is valuable across domains. Numeric overlap and unsupported-number detection should be added early in finance, followed by row/header and period alignment. Semantic evaluator metrics should remain in the

model because they capture context that deterministic rules miss. This layered design is easier to inspect than a single opaque score and can attach a reason to a low predicted probability.

The remaining error pattern points toward richer table semantics. A value can be present in the evidence and still be wrong for the question because it belongs to another period, subtotal, or denominator. Future table encoders should represent cell coordinates, hierarchical headers, footnote links, and arithmetic provenance. Those additions can extend the current feature family while preserving the central design principle: answer-quality prediction should make the evidence relationship explicit.

5. Limitations

The study evaluates the quality of answers already contained in RAGBench; it does not train a retriever or generator and does not claim that the hybrid model produces better answers. Its role is to estimate support and quality after a response and its retrieved evidence are available. The quality target also inherits the benchmark annotation design. Although combining four dimensions provides a useful graded outcome, a different weighting could be preferable in a setting where numerical materiality or regulatory compliance dominates other criteria.

The table-text features are deterministic and lightweight, but they are not a full financial table parser. They do not reconstruct merged headers, multi-level statement structure, footnote references, or the complete arithmetic program underlying every answer. The error slices show that copied values associated with the wrong period or denominator remain difficult. A richer representation of cell coordinates and calculation provenance would be needed for those cases.

The out-of-domain comparison is intentionally compact. eManual has only 132 test rows and represents procedural customer support rather than a second financial institution or reporting regime. It is useful for separating general evidence-use signals from table-specific signals, but it cannot establish broad cross-industry generalization. Additional evaluations should include annual reports from different accounting regimes, earnings-call transcripts, regulatory filings, and internal finance policies.

Finally, the automatic evaluator features inherit the behavior of their underlying evaluators. The ablation study shows that the hybrid model does not depend on those scores alone, but evaluator updates or generator changes can shift their distribution. A production deployment should monitor calibration over time and periodically validate threshold policies against expert review decisions.

6. Conclusion

This study presented answer-quality regression and adherence classification for financial RAG on RAGBench FinQA, with eManual as a procedural comparison domain. The hybrid model combines lexical information, automatic RAG evaluator scores, evidence-use counts, and deterministic table-text evidence features. It achieves 0.174 RMSE and 0.819 AUROC on FinQA, outperforming text-only, metric-only, evidence-only, and table-text-only alternatives. On eManual, it achieves 0.160 RMSE and 0.842 AUROC.

The results show that financial answer evaluation benefits from explicit checks for numeric support, units, reporting periods, table cues, and utilized evidence keys. These checks complement semantic evaluators rather than replace them. Calibration and error-slice analyses further show that the combined score can support review routing while exposing the reason for a low-confidence decision. A layered evaluator that joins domain-general evidence utilization with finance-specific table and numeric validation offers a practical foundation for monitoring grounded financial question-answering systems.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 9459-9474.

[2] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, “Retrieval-Augmented Generation for Large Language Models: A Survey,” arXiv:2312.10997, 2023.

[3] R. Friel, M. Belyi, and A. Sanyal, “RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems,” arXiv:2407.11005, 2024.

[4] Z. Chen, W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T. Huang, B. Routledge, and W. Y. Wang, “FinQA: A Dataset of Numerical Reasoning over Financial Data,” in Proc. 2021 Conf. Empirical Methods in Natural Language Processing, 2021, pp. 3697-3711, doi: 10.18653/v1/2021.emnlp-main.300.

[5] A. Nandy, S. Sharma, S. Maddhashiya, K. Sachdeva, P. Goyal, and N. Ganguly, “Question Answering over Electronic Devices: A New Benchmark Dataset and a Multi-Task Learning Based QA Framework,” in Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 4600-4609, doi: 10.18653/v1/2021.findings-emnlp.392.

[6] S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333-389, 2009.

[7] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, “Dense Passage Retrieval for Open-Domain Question Answering,” in Proc. 2020 Conf. Empirical Methods in Natural Language Processing, 2020, pp. 6769-6781.

[8] R. Nogueira and K. Cho, “Passage Re-ranking with BERT,” arXiv:1901.04085, 2019.

[9] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, “REALM: Retrieval-Augmented Language Model Pre-Training,” in Proc. 37th Int. Conf. Machine Learning, 2020, pp. 3929-3938.

[10] G. Izacard and E. Grave, “Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering,” in Proc. 16th Conf. European Chapter of the Association for Computational Linguistics, 2021, pp. 874-880.

[11] 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 preprint arXiv:2511.19481, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2511.19481

[12] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, “Lost in the Middle: How Language Models Use Long Contexts,” Transactions of the Association for Computational Linguistics, vol. 12, pp. 157-173, 2024.

[13] K. Zhang, S. Meng, and E. Zhou, “Evidence-Grounded Trading Desk Risk Memos over SEC Filings: Retrieval-Augmented Generation with XBRL Numeric Verification,” Journal of Applied Computing and Statistics, vol. 3, no. 2, pp. 60-76, Feb. 2023, doi: 10.69987/JACS.2023.30205.

[14] Q. Wu, J. Bai, and X. Zhou, “Evidence-Grounded Financial RAG: Reducing Numerical Hallucination in LLM-Generated Corporate Risk Memos,” Journal of Applied Computing and Statistics, vol. 3, no. 3, pp. 65-84, Mar. 2023, doi: 10.69987/JACS.2023.30306.

[15] S. Zhou, Z. Li, and E. Wang, “Evidence-Grounded RAG for Tokenized Trade Receivable Disclosure QA under U.S. Capital Market Standards,” Journal of Applied Computing and Statistics, vol. 3, no. 7, pp. 41-57, Jul. 2023, doi: 10.69987/JACS.2023.30704.

[16] S. Zhou, Z. Li, and E. Wang, “Long-Document RAG for Contractual and Insurance Clause Analysis in Receivables RWA Structures,” Journal of Applied Computing and Statistics, vol. 4, no. 8, pp. 88-104, Aug. 2024, doi: 10.69987/JACS.2024.40810.

[17] C. Li, J. Bai, and S. Wang, “Evidence-Chain Reliable RAG: Word-Level Hallucination Detection, Source Attribution, and Provenance Explanation for LLM Applications,” Journal of Applied Computing and Statistics, vol. 4, no. 2, pp. 76-92, Feb. 2024, doi: 10.69987/JACS.2024.40207.

[18] C. Li, W. Su, and E. Zhang, “Lightweight Hallucination Firewall for Enterprise LLM Applications: Evidence Consistency, Self-Checking, and Small-Model Detection on TruthfulQA,” Journal of Applied Computing and Statistics, vol. 3, no. 1, pp. 49-65, Jan. 2023, doi: 10.69987/JACS.2023.30104.

[19] B. Zhang, H. Rao, and D. Zhao, “Evidence-Grounded RAG for Cloud-Native DevOps: Hallucination-Resistant AIOps Question Answering over Private Operations Documents,” Journal of Applied Computing and Statistics, vol. 4, no. 3, pp. 109-125, Mar. 2024, doi: 10.69987/JACS.2024.40308.

[20] 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,” Journal of Applied Computing and Statistics, vol. 4, no. 4, pp. 97-111, Apr. 2024, doi: 10.69987/JACS.2024.40408.

[21] J. Nie and D. Zheng, “Ambiguity-Aware HDFS Log Anomaly Detection with Retrieval-Augmented Failure Narratives and Selective Refusal,” Journal of Applied Computing and Statistics, vol. 3, no. 1, pp. 66-80, Jan. 2023, doi: 10.69987/JACS.2023.30105.

[22] Y. Li, “Execution-Feedback and Retrieval-Augmented Generation for Conversational Text-to-SQL: From One-Shot Questions to Clarification-Driven Executable Dialogs,” Journal of Applied Computing and Statistics, vol. 3, no. 2, pp. 1-17, Feb. 2023, doi: 10.69987/JACS.2023.30201.

[23] Y. Li, “Test-in-the-loop LLM Repair: Verifiable Automated Program Repair on QuixBugs with a ‘Failing Test to Patch to Regression Test’ Loop,” Journal of Applied Computing and Statistics, vol. 4, no. 2, pp. 62-75, Feb. 2024, doi: 10.69987/JACS.2024.40206.

[24] Z. Li, K. Zhang, and A. Wong, “Numerical-Reasoning Guardrails for a Quant Research Assistant: A Compact Reproducible Benchmark Using SEC and FRED Data,” Journal of Technology Informatics and Engineering, vol. 5, no. 2, pp. 75-90, Jun. 2026, doi: 10.51903/jtie.v5i2.541.

[25] S. Zhou, Y. Chen, and K. Lee, “Accounting-Aware Evidence-Constrained Agents for Disclosure, Settlement, and Secondary-Market Risk Monitoring in Tokenized,” Journal of Technology Informatics and Engineering, vol. 5, no. 2, pp. 60-74, Jun. 2026, doi: 10.51903/jtie.v5i2.544.

[26] 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,” Journal of Applied Computing and Statistics, vol. 3, no. 4, pp. 31-47, Apr. 2023, doi: 10.69987/JACS.2023.30403.

[27] 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,” Journal of Applied Computing and Statistics, vol. 4, no. 6, pp. 74-85, Jun. 2024, doi: 10.69987/JACS.2024.40606.

[28] G. Salton and C. Buckley, “Term-Weighting Approaches in Automatic Text Retrieval,” Information Processing & Management, vol. 24, no. 5, pp. 513-523, 1988.

[29] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[31] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall, 1993.

[32] G. W. Brier, “Verification of Forecasts Expressed in Terms of Probability,” Monthly Weather Review, vol. 78, no. 1, pp. 1-3, 1950.

[33] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” in Proc. 34th Int. Conf. Machine Learning, 2017, pp. 1321-1330.