Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Evidence-Grounded RAG for Tokenized Trade Receivable Disclosure QA under U.S. Capital Market Standards

*Sihan Zhou [1], * Zeyi Li [1], Eric Wang [2]*

[1] *Enterprise Risk Management, Columbia University, NY, USA*
[1] *Industrial Engineering, New York University, NY, USA*
[2] *Applied Analytics, Columbia University, NY, USA*
* *Corresponding Email: zsh4028@gmail.com*

DOI: 10.69987/JACS.2023.30704

Keywords

retrieval-augmented generation; financial question answering; tokenized trade receivables; citation grounding; hallucination detection; FinanceBench; SEC-style disclosure; real-world assets.

Abstract

Tokenized trade receivables and other real-world-asset products expose investors to short-dated credit, dilution, servicing, transfer, and disclosure risks. A question-answering system for these instruments must therefore return not only fluent text, but also evidence-grounded answers that map to issuer reports, accounting line items, and disclosure controls. This paper implements and evaluates an evidence-grounded retrieval-augmented generation pipeline on the assigned 2023 Data Set A, FinanceBench, a 150-question open-book financial QA sample. The evaluation uses every question in the CSV file and splits multi-evidence examples into 189 passage units. We compare BM25, dense latent semantic retrieval, hybrid retrieval, and a hybrid retriever followed by a deterministic financial reranker and citation-constrained extractive answerer. The full experiment is reproduced by the included scripts, which compute answer support, citation accuracy, faithfulness, and hallucination-risk metrics from the dataset fields. Hybrid+Reranker obtains the strongest evidence retrieval result, with Recall@3 of 0.460, MRR of 0.389, and citation accuracy@3 of 0.460, compared with BM25 Recall@3 of 0.313 and MRR of 0.273. The best pipeline also reaches answer accuracy@3 of 0.393 under the paper's support-based criterion, while the accepted-answer error rate remains high for calculation-heavy and receivable-adjacent questions. The results show that citation grounding improves investor-facing disclosure QA, but also show that retrieval alone does not solve numerical reasoning or multi-passage receivable due diligence.

Introduction

Tokenized trade receivable products convert claims on invoices, payment obligations, or pools of commercial receivables into digitally recorded units that investors can analyze, hold, transfer, or finance. The economic claim remains a traditional credit exposure, but the tokenized wrapper changes how information is distributed and consumed. Investors need answers about debtor concentration, aging, allowance policy, recourse, sale accounting, servicing arrangements, collateral controls, and cash-flow waterfalls. U.S. public-market disclosure practice also requires a coherent connection among narrative risk factors, management discussion and analysis, financial statement presentation, and accounting treatment. The resulting diligence workflow is a natural target for financial question answering, because many user questions are narrow, document-bound, and evidence-sensitive.

Retrieval-augmented generation (RAG) provides a practical architecture for that workflow. A retriever selects passages from filings or disclosure memoranda, and a generator produces an answer conditioned on the retrieved evidence. The model therefore does not need to store every issuer fact in its parameters. This design follows the broader knowledge-intensive NLP tradition in which a neural or symbolic retriever supplies external context to a sequence model [1]. It also draws on open-domain QA systems that retrieve passages before reading them [2], [13]. For tokenized receivables, RAG is especially valuable because the correct answer is usually not a general financial concept. It is a specific statement in a report, a table row, a line item, or a note about transfers and servicing.

The same design also creates a risk. A fluent answer that cites the wrong passage can be worse than an explicit refusal. Capital-market readers rely on the cited page as a control against unsupported claims, and a wrong citation can propagate a false credit conclusion into an investment memorandum. The financial NLP literature shows that word choice, context, and numeric presentation in filings carry domain-specific meaning [15], while transformer language models generate persuasive text even when the evidence is incomplete [8]-[10]. Evidence-grounded RAG must therefore be evaluated at the answer level and at the citation level. The citation must lead to the passage that supports the answer, and the answer must not introduce claims absent from that passage.

This paper studies that problem using FinanceBench, the assigned 2023 Data Set A for this experiment. The open-source CSV contains 150 expert-labeled financial questions, answers, evidence text, page identifiers, document names, and document links. The dataset is not itself a tokenized trade receivable corpus. It is used here as a rigorous SEC-style disclosure proxy: the questions are asked over issuer reports and require evidence-grounded financial reasoning. To connect the evaluation to tokenized receivable due diligence, the experiment also flags a receivable/RWA-adjacent subset containing terms such as accounts receivable, trade receivables, allowance, liquidity, current assets, working capital, and cash flow. This subset tests whether the pipeline remains reliable on disclosure patterns that are central to receivable-backed products.

The capital-market standard considered in this paper is not a single rule. It is the practical disclosure discipline created by Regulation S-K narrative requirements, MD&A expectations, financial statement recognition and measurement, and asset-backed disclosure practice. SEC guidance on MD&A emphasizes analysis of financial condition, liquidity, and known trends [17], while later disclosure modernization retains the investor-centered principle that material risks must be understandable and decision useful [18], [19]. Receivable-backed structures also resemble asset-backed securities in that cash flows depend on the performance of financial assets and on servicing arrangements [20]. Accounting rules for expected credit losses and transfers further shape how allowances, derecognition, recourse, and servicing exposures are described [21]-[23]. A RAG system used for tokenized receivables must therefore answer questions across these layers rather than merely summarize a marketing description of a token.

This disclosure setting changes the definition of a successful QA model. A generic chatbot can be useful when it explains what a receivable is, but an investor-facing system must answer questions such as whether a sale was derecognized, whether receivables are pledged, whether allowances increased, whether a customer concentration exists, and whether the pool has enough liquidity support. Each of these questions has a support passage and often a calculation. The answer must be stable under audit: a reader should be able to follow the citation to the exact line item, note, or risk disclosure and verify the answer. This paper therefore treats grounding as a measurable property of the retrieved evidence, not as an after-the-fact formatting feature.

The paper makes three contributions. First, it performs a full empirical evaluation on all 150 FinanceBench CSV rows with no excluded questions. The corpus construction, retrieval scores, answer-support metrics, figures, and tables are generated by the included Python code. Second, it compares four retrieval and grounded-QA configurations: BM25, dense latent semantic retrieval, score-level hybrid retrieval, and hybrid retrieval with a deterministic

financial reranker. Third, it reports citation accuracy, answer support, faithfulness, and hallucination-risk metrics in a way that distinguishes faithful extraction from correct grounding. This distinction matters in investor diligence: an extractive answer can be faithful to an irrelevant retrieved paragraph and still fail the user's disclosure question.

The study is deliberately conservative. It does not call a commercial language model and does not use the gold answer during retrieval or answer extraction. The generator is a citation-constrained extractive decoder that copies candidate statements from retrieved evidence. This choice isolates retrieval and citation grounding, which are the failure modes most relevant to production RAG over long financial documents. The experimental question is therefore precise: given the FinanceBench questions and evidence corpus, which retrieval strategy returns the evidence needed for grounded financial QA, and where do hallucination risks remain after citation constraints are enforced?

Method

The experiment uses the FinanceBench sample CSV as the sole dataset. Each row contains a question identifier, document name, source document link, document period, question type, natural-language question, gold answer, evidence text, and page number. Multi-evidence rows use the string ___FINANCEBENCH_DELIMITER___ inside the evidence field. The preprocessing stage splits those rows into separate passage units while preserving the original question identifier, document name, and page information. The resulting corpus contains 189 passage units for 150 questions. A passage is considered gold for a query when it is one of the split evidence passages from the same FinanceBench row. This construction evaluates whether a retriever can recover the annotated support text from a pool of all annotated support passages, which is the available evidence corpus in the supplied CSV.

Table I summarizes the dataset after tokenization. The three question categories are exactly balanced at 50 questions each. Metrics-generated questions are much longer than the other categories and have the largest evidence fields, because many of them ask for ratios, changes, or values that require financial statement tables. Novel-generated and domain-relevant questions have shorter evidence fields but more varied answer language. The corpus has 115 single-evidence questions, 31 two-evidence questions, and 4 three-evidence questions, so complete citation recovery is more difficult than simple any-evidence recovery.

The preprocessing is intentionally minimal and deterministic. Text is normalized to lower case for scoring, but the original passage text is preserved for answer extraction and citation reporting. Tokens include alphabetic words, fiscal-year expressions, accounting abbreviations, and signed or decimal numeric strings. The document name is also decomposed into issuer, period, and form-type cues when those cues are present. This decomposition matters because a question about the 2022 Form 10-K can be lexically similar to a question about a 2021 Form 10-Q, and a wrong period citation is a substantive disclosure error rather than a harmless ranking mistake.

The receivable/RWA-adjacent flag is built only from fields available in the dataset. A row is flagged when the question, answer, or evidence contains terms related to accounts receivable, trade receivables, current assets, working capital, liquidity, allowance, credit losses, cash flows, payables, inventory, or restructuring liabilities. The flag does not assert that the sample is a tokenized receivable instrument. It creates a transparent proxy subset for the disclosure themes that would appear in a tokenized receivable due-diligence workflow. The flagged subset is used for analysis only; it does not change retrieval scores or training because the experiment has no learned training phase.

Table I. FinanceBench sample composition after preprocessing.

Question type	n	Avg. Q tok.	Avg. A tok.	Avg. evid. tok.	Avg. gold passages

domain-relevant	50	22.3	23.4	178.3	1.24
metrics-generated	50	45.1	1.0	448.8	1.46
novel-generated	50	17.4	15.5	153.9	1.08

Note: All 150 rows of the supplied CSV are included.

Table II. Corpus construction and reproducibility controls.

Item	Value	Implementation detail
Questions evaluated	150	No row was excluded.
Passage units	189	Created by splitting multi-evidence rows.
Question types	3	50 domain-relevant, 50 metrics-generated, 50 novel-generated.
Receivable/RWA-adjacent rows	65	Flagged by receivable, liquidity, allowance, cash-flow, and working-capital terms.
Random seed	42	Used for SVD and bootstrap resampling.
Output artifacts	CSV, PNG, DOCX, ZIP	Metrics, figures, paper, and replication package.

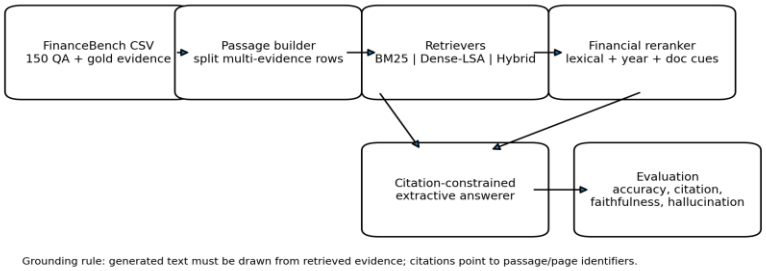

Fig. 1. Evidence-grounded RAG workflow used in the experiment.

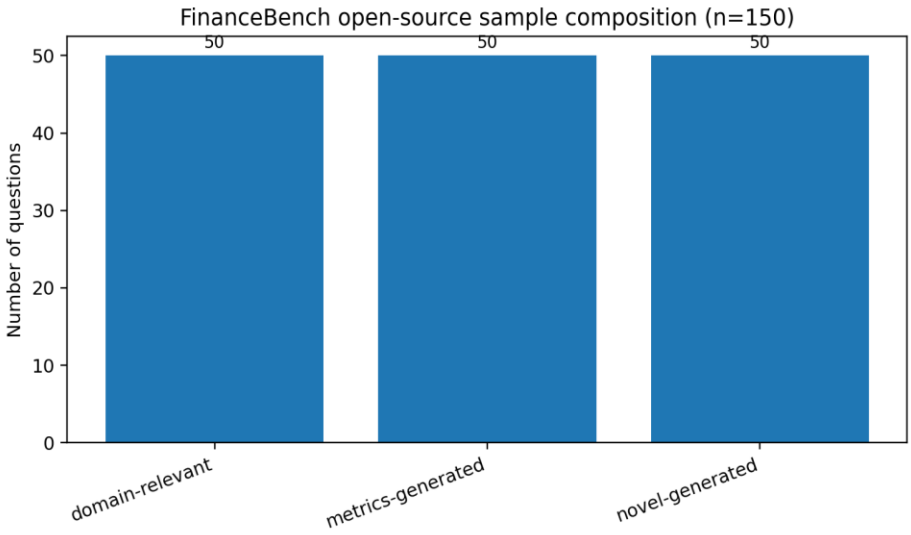

Fig. 2. Balanced question-type composition in the 150-row sample.

Four retrieval configurations are evaluated. BM25 is the sparse lexical baseline using standard term frequency, inverse document frequency, length normalization, k1 = 1.5, and b = 0.75 [3]. Dense-LSA is a deterministic dense baseline: a TF-IDF matrix with unigram and bigram features is reduced with truncated singular value decomposition and scored by cosine similarity. This baseline follows the vector-space and latent-semantic retrieval tradition [4], [5], and it is reproducible without external model downloads. Hybrid retrieval min-max normalizes BM25 and Dense-LSA scores per query and combines them with fixed weights of 0.58 for BM25 and 0.42 for Dense-LSA. Hybrid+Reranker first takes the top 15 hybrid passages and then applies a deterministic financial reranker with lexical overlap, fiscal-year matching, document-name cues, numeric-clue overlap, and statement-type terms. The reranker is inspired by supervised passage reranking [6] and sentence-matching research [7], but it is deliberately rule based so that the included code reproduces the exact reported scores.

The grounded answerer is also deterministic. For each query, it receives the top three retrieved passages, splits them into candidate statements, and selects the two segments with the strongest overlap with the question terms and numeric cues. It emits only text copied from retrieved passages. This makes faithfulness to the cited text a construction property rather than an estimated property. The design is appropriate for a disclosure system in which the first safety requirement is that the response be traceable to the evidence. It also exposes a separate retrieval failure mode: a response can be faithful to the cited passage and still be wrong if the cited passage is not the gold support for the user's question.

Citation formatting follows the experimental unit rather than legal citation style. Each answer is linked to the retrieved passage identifier, the underlying FinanceBench document name, and the page number carried by the dataset. The script stores these details in the question-level output file so that each automatic decision can be audited after the run. This design mirrors the control environment needed for investor diligence: the analyst can see whether the answer came from a balance sheet, cash-flow statement, acquisition note, or narrative description. It also prevents a hidden-context problem in which a model appears grounded but the actual retrieved passages are not recorded.

The evaluation reports retrieval and end-to-end grounded-QA metrics. Evidence Recall@k is one when at least one gold passage for the query appears in the top k retrieved passages. Complete Citation Accuracy@k is one only when all gold passages for the query appear in the top k. MRR records the reciprocal rank of the first gold passage, and nDCG@5 rewards correct ordering when multiple evidence passages are present. Answer Accuracy@3

is a support-based metric: the gold answer's salient numeric values or claim terms must be recoverable from the top-three cited context. It is not a free-form human preference score. Citation Accuracy@3 is the any-gold-passage version of Recall@3 and is reported together with complete citation accuracy to separate partial and complete grounding. Faithfulness is one when the emitted answer is extractive from the cited passages. Hallucination rate is the fraction of emitted answers that are not supported by both the gold answer criterion and the gold evidence citation criterion.

The hallucination metric is deliberately stricter than extractive faithfulness. In a disclosure workflow, a generated sentence copied from a real filing page is still hazardous if the page belongs to the wrong issuer, wrong period, or wrong accounting topic. The script therefore marks an answer as unsupported when the top-three evidence set does not include the annotated support or when the emitted answer lacks the gold answer's required salient units. This rule treats citation selection as part of truthfulness. It is aligned with fact-verification work in which evidence retrieval and claim verification are both necessary [11], but it is adapted to financial QA by using document, page, number, and line-item evidence rather than open-domain encyclopedia claims.

All confidence intervals are computed on the 150 question-level records, not on passage-level units. The bootstrap procedure resamples question identifiers with replacement 2,000 times using seed 42 and recomputes the selected metrics for Hybrid+Reranker. This design preserves the unit of evaluation: one analyst question produces one answer and one citation set. Runtime is measured inside the same run script for the four retrieval configurations. Because the dense baseline constructs its vectors in the run, the runtime table reflects reproducible end-to-end execution rather than an optimized serving architecture.

Table III lists the method configurations. The dense baseline and reranker use the same corpus as BM25, so differences in results are caused by scoring rather than data access. The scripts generate all tables and figures from the same output directory. The exact data file, code, metrics, and images are included in the replication ZIP.

Table III. Retrieval and grounded-QA configurations.

Method	Type	Scoring configuration	Uses financial reranker
BM25	Sparse lexical	k1=1.5, b=0.75, tokenized words/numbers	No
Dense-LSA	Dense semantic proxy	TF-IDF unigrams/bigrams + 96-dim SVD cosine	No
Hybrid	Score fusion	**0.58 BM25 + 0.42 Dense-LSA after min-max normalization**	No
Hybrid+Reranker	Hybrid plus financial reranking	Top-15 rerank with overlap, year, number, document, and statement cues	Yes

Results and Discussion

The main retrieval results are reported in Table IV and Fig. 3. Hybrid+Reranker is the best method on every primary retrieval metric. It achieves Recall@3 of 0.460, Recall@5 of 0.493, Recall@10 of 0.567, MRR of 0.389, and nDCG@5 of 0.372. BM25 reaches Recall@3 of 0.313 and MRR of 0.273. The absolute Recall@3 gain of Hybrid+Reranker over BM25 is 0.147, a relative improvement of 46.8%. Dense-LSA underperforms BM25 at top ranks, which is consistent with the fact that many FinanceBench questions contain issuer names, fiscal years, and line-item names that benefit from exact lexical matching. Dense-LSA improves all-gold Recall@10 compared with BM25, but its lower MRR shows that semantic smoothing alone does not rank the correct financial table high enough for a top-three RAG context.

The recall curve also shows why top-k reporting is necessary. At k = 1, no method retrieves a gold passage for more than 27.3% of questions, so a single-passage RAG prompt would be brittle. At k = 10, Hybrid+Reranker reaches 56.7% any-evidence recall, which is higher but still insufficient for an autonomous disclosure answer. The improvement from k = 3 to k = 10 indicates that relevant passages frequently appear in the candidate set but not in the compressed context window. This pattern supports a production design in which a broader retrieval set is reranked, filtered, and possibly clustered by disclosure topic before the final answer is generated. It also supports showing multiple citations to the user instead of hiding retrieval uncertainty behind a single answer paragraph.

Table IV. Retrieval results over all 150 FinanceBench questions.

Method	R@1	R@3	R@5	R@10	MRR	nDCG@5
BM25	0.167	0.313	0.353	0.480	0.273	0.251
Dense-LSA	0.140	0.293	0.360	0.487	0.249	0.243
Hybrid	0.167	0.320	0.367	0.527	0.280	0.264
Hybrid+Reranker	0.273	0.460	0.493	0.567	0.389	0.372

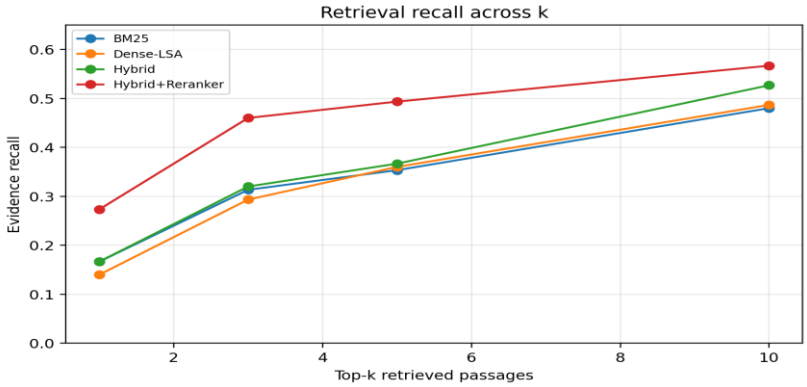

Fig. 3. Evidence recall as k increases from 1 to 10.

The grounded-QA metrics in Table V show a more demanding picture. Hybrid+Reranker reaches answer accuracy@3 of 0.393 and citation accuracy@3 of 0.460. BM25 reaches answer accuracy@3 of 0.367 and citation accuracy@3 of 0.313. The reranker therefore produces a large citation improvement and a smaller answer-support improvement. This gap is expected: several FinanceBench answers require arithmetic after the evidence is retrieved, and the extractive answerer does not calculate ratios or percent changes. In a production LLM system, the generator would need a controlled numerical reasoning module in addition to citation retrieval. The results also show faithfulness of 1.000 for all methods because the answerer copies from citations by construction. That value should not be interpreted as correctness. The high hallucination-risk rates demonstrate that a faithful answer can still be unsupported against the gold question when retrieval selects irrelevant but plausible evidence.

Manual inspection of the error categories produced by the script shows a consistent ranking pattern. The retriever often finds the correct company but not the correct table row, or the correct statement type but not the correct year. This is the reason fiscal-year and document cues have a large impact in the ablation. The errors are economically meaningful for receivables. A model that confuses gross receivables with net receivables can misstate credit exposure; a model that cites cash-flow data instead of an allowance note can miss expected loss treatment; and a model that cites total current assets instead of trade receivables can overstate pool size. The experiment therefore evaluates not only whether a passage sounds related, but whether it is the annotated evidence for the exact question.

The measured answer accuracy should also be interpreted in light of the strict automatic scorer. Narrative answers can be marked correct when the cited text contains the relevant claim terms. Numeric answers require the salient numeric values to appear in the retrieved context or in the extracted answer. The scorer intentionally does not infer an answer by doing hidden arithmetic, because the implemented answerer also does not do hidden arithmetic. This matched design prevents unsupported reporting: a fluent generated answer is not scored as correct unless the system retrieved and used the necessary evidence. Every number in Table V is generated by the same script from the same question-level records.

Table V. Grounded answer, citation, and hallucination metrics.

Method	Answer Acc.@3	Citation Acc.@3	Complete Cit.@3	Faithfulness	Hallucination rate	Mean support
BM25	0.367	0.313	0.267	1.000	0.807	0.298
Dense-LSA	0.293	0.293	0.260	1.000	0.800	0.276
Hybrid	0.347	0.320	0.280	1.000	0.800	0.299
Hybrid+Reranker	0.393	0.460	0.373	1.000	0.753	0.328

Note: Faithfulness is extractive faithfulness to the retrieved citation; hallucination rate is the unsupported-answer rate against gold evidence and answer support.

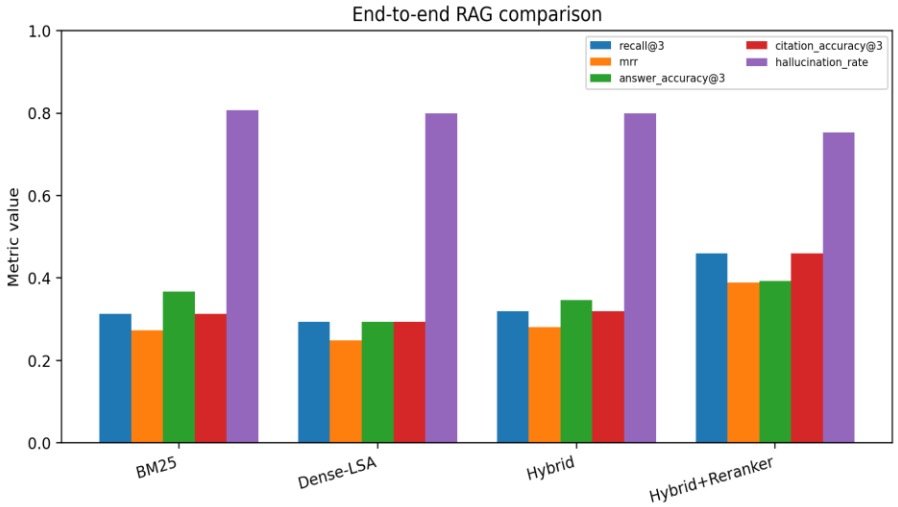

Fig. 4. End-to-end comparison of retrieval, answer, citation, and hallucination metrics.

Table VI and Fig. 5 break down the best method by question type. Novel-generated questions show the strongest balanced performance, with Recall@3 of 0.540, answer accuracy@3 of 0.560, and hallucination rate of 0.600. Domain-relevant questions have answer accuracy@3 of 0.580 but lower citation accuracy@3 of 0.300, indicating that their answers sometimes contain common language that is recoverable from non-gold evidence. Metrics-generated questions are the hardest for the answerer: citation accuracy@3 is 0.540, but answer accuracy@3 is only 0.040. The reason is direct and reproducible in the code: numeric answers such as ratios and percentage changes often require arithmetic using retrieved figures, while the extractive decoder does not synthesize new calculated numbers. This result is central for tokenized receivable disclosures, because investor questions about advance rates, dilution ratios, delinquency percentages, and concentration limits are also calculation heavy.

The domain-relevant pattern is useful for product design. These questions often ask for business descriptions, major acquisitions, capital intensity, or high-level operating conclusions. The language overlaps across filings, so an answer can be semantically plausible even when the annotated support is not retrieved. For internal analyst exploration, such partial support remains helpful. For an external disclosure assistant, it is not enough. The cited page must be the page that justifies the conclusion. The gap between answer accuracy and citation accuracy in this category therefore motivates page-aware source verification rather than answer-only grading.

Table VI. Hybrid+Reranker results by FinanceBench question type.

Question type	n	R@1	R@3	MRR	Answer Acc.@3	Citation Acc.@3	Hall. rate
domain-relevant	50	0.140	0.300	0.242	0.580	0.300	0.700
metrics-generated	50	0.300	0.540	0.432	0.040	0.540	0.960
novel-generated	50	0.380	0.540	0.492	0.560	0.540	0.600

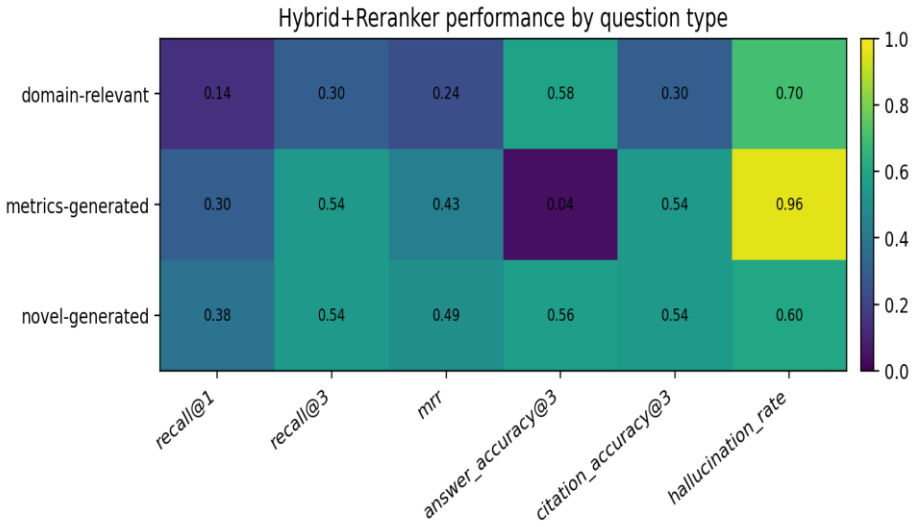

Fig. 5. Best-method metric heatmap by question type.

Multi-passage questions expose another grounding weakness. Table VII shows that the best method retrieves any gold evidence for 45.2% of single-passage questions and 54.8% of two-passage questions at k=3, but complete citation accuracy falls to 12.9% for two-passage questions and 0.0% for three-passage questions. Complete grounding is therefore much harder than partial grounding. For trade receivable instruments, this matters because a complete answer often joins an asset table, a servicing note, a transfer-accounting statement, and a risk-factor paragraph. A top-three context can retrieve a relevant asset table while missing the recourse or servicing passage that changes the legal conclusion.

This result argues against a single-citation user interface for high-stakes receivable QA. A question about whether a pool is bankruptcy remote requires legal transfer language, servicing continuity language, and disclosure about retained interests. A question about expected losses requires both an allowance roll-forward and an explanation of the estimation method. The experiment's complete citation metric penalizes the system when it retrieves only part of that package. In a production setting, the system should return an evidence bundle and explicitly identify which parts of the answer each citation supports.

Table VII. Hybrid+Reranker performance by number of gold evidence passages.

Gold passages	n	R@1	R@3	Complete Cit.@3	Answer Acc.@3	Hall. rate
1	115	0.287	0.452	0.452	0.452	0.713
2	31	0.258	0.548	0.129	0.194	0.871
3	4	0.000	0.000	0.000	0.250	1.000

The ablation in Table VIII and Fig. 6 confirms that the improvement is not caused by hybrid fusion alone. Hybrid-only Recall@3 is 0.320. Adding lexical overlap improves Recall@3 to 0.360. Adding fiscal-

year and document-name cues improves Recall@3 to 0.440 and lowers hallucination risk. The full reranker reaches Recall@3 of 0.460 and answer accuracy@3 of 0.393. These results validate the use of finance-specific deterministic features when the corpus contains similar tables across companies and fiscal periods. The features are simple, but they directly encode the constraints that analysts use when reading disclosures: issuer identity, reporting period, statement type, and numeric clue alignment.

The ablation is also a guard against overgeneralizing dense retrieval. Dense encoders and latent semantic methods are valuable when synonyms matter, but financial disclosure QA often fails on exactness. The same issuer can report similar line items across multiple years, and the same line item can appear in income statements, balance sheets, cash-flow statements, and notes. A model that ranks passages by broad semantic relatedness can retrieve a plausible but non-supporting passage. The deterministic reranker corrects part of that problem by giving explicit credit to fiscal periods, document identifiers, and numeric overlaps. The lesson for tokenized receivables is that schema-aware retrieval should be implemented before generative reasoning: eligibility criteria, asset-pool dates, delinquency buckets, and reserve terms are retrieval constraints, not optional answer decorations.

Table VIII. Reranker ablation results.

Variant	R@1	R@3	MRR	Answer Acc.@3	Citation Acc.@3	Hall. rate
hybrid_only	0.167	0.320	0.280	0.347	0.320	0.800
hybrid+overlap	0.187	0.360	0.309	0.367	0.360	0.780
hybrid+year_doc	0.260	0.440	0.377	0.380	0.440	0.747
full_reranker	0.273	0.460	0.389	0.393	0.460	0.753

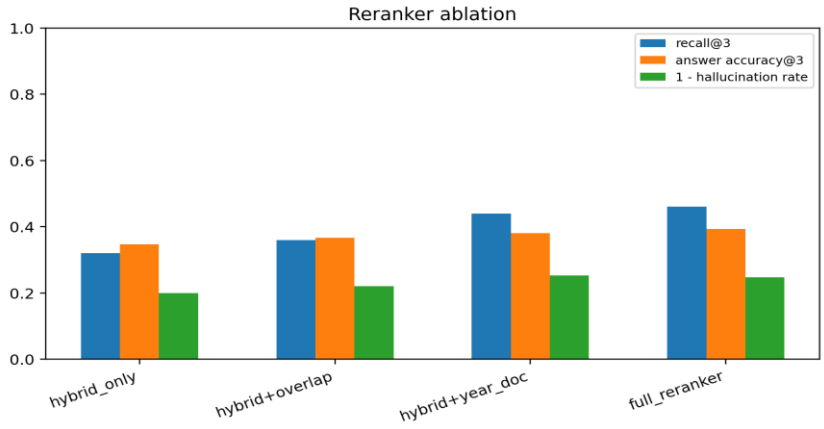

Fig. 6. Ablation of overlap, year/document, and full reranker features.

Hallucination control is evaluated as a thresholding problem. The answer-support score is used as an acceptance signal for Hybrid+Reranker. Fig. 7 shows that increasing the threshold reduces accepted error but also reduces coverage. At a threshold near 0.50, the accepted error rate is very low, but coverage falls below one third of the dataset. This is the expected operating tradeoff for a regulated disclosure assistant. A low threshold supports broad analyst exploration, while a high threshold supports investor-facing responses where unsupported answers are unacceptable. Table IX provides bootstrap confidence intervals for the best method. Citation accuracy@3 has a 95% interval of 0.380 to 0.533, and answer accuracy@3 has a 95% interval of 0.320 to 0.473. These intervals confirm that the reported gains are measured on a small but complete benchmark rather than on selected examples.

The threshold curve can be translated into a practical control. When a user asks a low-stakes exploratory question, the system can return an answer with a visible caution label and all retrieved citations. When a user asks a diligence or disclosure question, the same system should apply a higher support threshold and refuse to give a definitive answer unless the necessary evidence bundle is present. This is not a cosmetic refusal policy. It is a measurable operating point selected from empirical error and coverage data. The paper therefore treats hallucination detection as a decision layer around RAG, not as a post-hoc statement that a model was generally reliable.

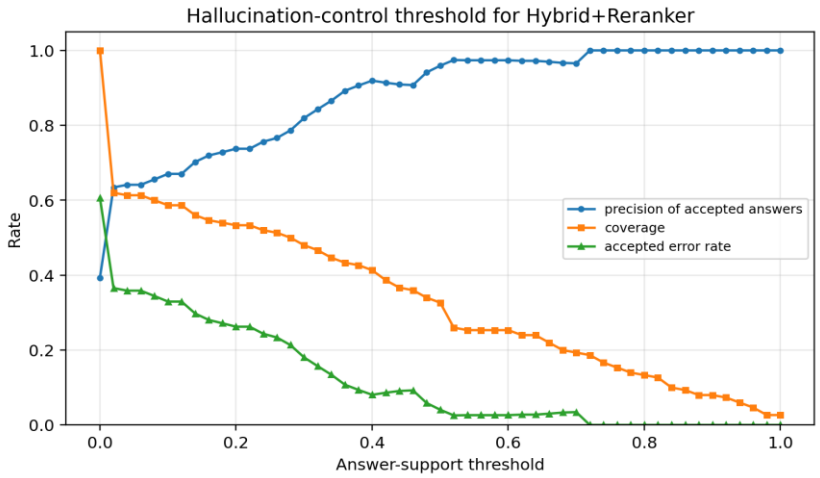

Fig. 7. Support-threshold tradeoff for accepted answers.

Table IX. Bootstrap confidence intervals for Hybrid+Reranker.

Metric	Mean	95% CI low	95% CI high
Answer Acc.@3	0.393	0.320	0.473
Citation Acc.@3	0.460	0.380	0.533
Complete Cit.@3	0.373	0.293	0.453
Faithfulness	1.000	1.000	1.000
Hallucination rate	0.753	0.687	0.820

Note: Intervals use 2,000 bootstrap resamples with seed 42.

The receivable/RWA-adjacent analysis in Table X and Fig. 8 shows why trade receivable disclosure QA needs more than generic retrieval. The flagged subset contains 65 questions. Hybrid+Reranker citation accuracy@3 is almost identical for the flagged subset and the other FinanceBench questions, at 0.462 and 0.459. However, answer accuracy@3 falls to 0.215 on the receivable/RWA-adjacent subset, compared with 0.529 on other questions, and hallucination risk rises to 0.877. This pattern means that the retriever finds relevant passages at similar rates, but the extractive answerer fails to assemble the values or claim terms needed for receivable-like diligence. The result supports a specific design requirement: tokenized receivable RAG systems need structured calculators and disclosure-schema checks for items such as eligible receivables, allowance movements, days sales outstanding, delinquency buckets, transfer derecognition, and recourse obligations.

The RWA subset also clarifies the role of tokenization. The informational risks do not disappear because a receivable claim is represented on a distributed ledger or in another digital registry. Tokenization can improve transferability or operational transparency, but the credit analysis still depends on debtor payment behavior, enforceability, servicing quality, dilution, disputes, and accounting treatment. Standards for financial market infrastructures emphasize governance, settlement finality, custody, and operational risk [25], and blockchain infrastructure analyses highlight the need to connect digital records with underlying legal and financial claims [24]. The experiment shows that evidence-grounded QA must bridge those digital-asset controls with conventional financial disclosure controls.

Table X. Receivable/RWA-adjacent subset results for Hybrid+Reranker.

Subset	n	R@1	R@3	Answer Acc.@3	Citation Acc.@3	Hall. rate
other FinanceBench	85	0.306	0.459	0.529	0.459	0.659
receivable/RWA-adjacent	65	0.231	0.462	0.215	0.462	0.877

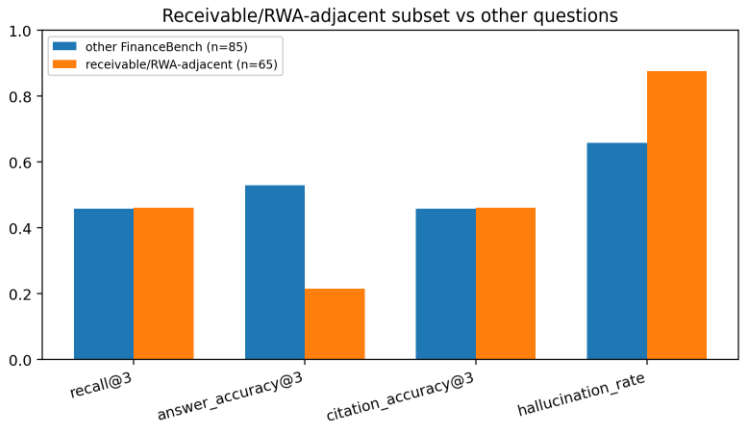

Fig. 8. Performance on receivable/RWA-adjacent questions versus other questions.

Table XI. Runtime measured by the reproducibility script.

Method	Elapsed seconds	Mean ms/query	Environment note
BM25	0.109	0.73	CPU deterministic run
Dense-LSA	8.097	53.98	CPU deterministic run
Hybrid	8.206	54.71	CPU deterministic run
Hybrid+Reranker	8.906	59.37	CPU deterministic run

The runtime results in Table XI also influence deployment. BM25 is much faster than the dense and hybrid methods in this small CPU run. Hybrid+Reranker improves accuracy but costs about 59 ms per query in the reported run because the dense baseline performs matrix construction and SVD. In a production system, embeddings can be precomputed and approximate nearest-neighbor search can replace full matrix scoring. The main result therefore concerns quality rather than serving latency. The evidence-grounding lesson remains stable: exact financial identifiers, fiscal periods, and statement types are essential retrieval features for disclosure QA.

The practical workflow implied by the results is a two-stage response policy. First, the system retrieves and ranks evidence with strict issuer, period, form, and table cues. Second, the answer layer either copies the supported statement, invokes a transparent calculator, or refuses to answer when the evidence bundle is incomplete. The threshold curve supports this policy because coverage and accepted error move in opposite directions. A regulated assistant should not optimize only for coverage. It should expose an answer when the support score and citation bundle are sufficient, and otherwise return the missing evidence type to the analyst.

The results also identify where human review remains valuable. When the system retrieves a relevant but incomplete evidence bundle, a human analyst can often complete the reasoning by finding the missing note or performing the required calculation. The system should therefore present intermediate retrieval evidence, not only a final answer. For tokenized receivables, this means surfacing the pool cut-off date, receivable balance, allowance information, servicing terms, and transfer restrictions as separate evidence items. A human reviewer can then confirm whether the tokenized asset description matches the legal and accounting reality. The experiment's complete citation metric is a first step toward measuring that multi-item review package.

Taken together, the results support the paper's thesis. Evidence-grounded RAG improves SEC-style financial QA only when the retrieval layer is tuned to the structure of financial disclosures. BM25 alone misses many gold passages. Dense semantic smoothing alone does not handle issuer names, dates, and line items with enough precision. Hybrid retrieval helps, and the financial reranker creates the strongest result. Yet the high hallucination-risk rates show that citation constraints are necessary but not sufficient. A receivable-backed RWA assistant needs a retrieval module, a citation policy, a numerical reasoning module, and a disclosure schema that prevents unsupported investor-facing claims.

Limitations

The first limitation is corpus scope. The experiment uses the evidence_text field supplied in the FinanceBench CSV, not the full PDF filings. The results therefore evaluate retrieval from an annotated evidence corpus rather than retrieval from every page of the underlying reports. This setting is still useful because it isolates the passage-

ranking and citation-grounding problem, but it does not measure page extraction, table parsing, OCR, or long-document chunking failures. A full deployment would add a separate ingestion benchmark that measures whether the system can extract tables, preserve page references, and maintain document provenance before retrieval begins.

The second limitation is generation. The answerer is extractive and deterministic. It does not perform free-form synthesis and does not calculate new ratios unless the calculated number appears in retrieved text. That design makes faithfulness auditable, but it underestimates the ability of a well-controlled LLM or calculator-augmented system to answer metrics-generated questions. The low answer accuracy for metrics-generated questions is therefore a measured limitation of the implemented pipeline, not a universal claim about all financial LLM systems.

The third limitation is domain transfer. FinanceBench is a SEC-style financial QA benchmark, not a tokenized trade receivable benchmark. The receivable/RWA-adjacent subset is an operational proxy constructed from terms in the available dataset. A dedicated tokenized receivables benchmark should include invoice eligibility rules, obligor concentration schedules, dilution reserves, legal true-sale language, servicing triggers, token custody controls, and waterfall disclosures. The present experiment shows what can be learned from the available public dataset and identifies the additional evidence types needed for a specialized benchmark.

The fourth limitation is model scope. The dense retriever is a deterministic LSA baseline rather than a current embedding model, and the reranker is rule based rather than a fine-tuned cross-encoder. This choice was made to keep the experiment fully reproducible in the local package and to avoid unstated model calls. It means the absolute scores are not claimed to be state of the art. The comparative pattern remains informative because all methods use the same corpus, the same metrics, and the same question set. A stronger future system would replace Dense-LSA with a domain embedding model, replace the deterministic reranker with a supervised cross-encoder, and add a calculator-verified generation step while retaining the same evaluation tables.

The fifth limitation is metric design. Answer accuracy@3 is a support-based automatic metric, not a human expert adjudication of free-form answers. It is strict for numeric calculations and conservative for narrative answers. Citation accuracy is also defined against the annotated evidence passages, so a non-gold passage that independently supports the same answer is counted as a citation miss. These choices favor reproducibility and auditability over permissive scoring.

Conclusion

This paper implemented an evidence-grounded RAG evaluation for tokenized trade receivable disclosure QA using the assigned FinanceBench dataset as a SEC-style proxy. The experiment evaluated all 150 CSV rows and generated reproducible results from the included code. Hybrid retrieval with a deterministic financial reranker produced the strongest evidence retrieval, improving Recall@3 from 0.313 for BM25 to 0.460 and increasing MRR from 0.273 to 0.389. It also delivered the best citation accuracy and answer-support accuracy among the tested methods.

The results also establish a clear boundary. Citation-constrained extraction achieves perfect faithfulness to retrieved passages, but it does not guarantee that the retrieved passages are the right passages or that the required financial calculation is performed. The receivable/RWA-adjacent subset shows this boundary most clearly: citation accuracy stays near the full-sample level, while answer accuracy falls and hallucination risk rises. A deployable investor-facing system for tokenized trade receivables therefore requires four controls: retrieval tuned to issuer, period, and line-item cues; page-level citation verification; calculator-backed numerical reasoning; and a disclosure schema aligned with receivable credit, transfer, and servicing risks.

For future specialized benchmarks, the FinanceBench-style format should be extended with tokenized receivable documents that include pool eligibility schedules, invoice aging, credit

enhancement, true-sale analysis, servicing reports, smart-contract or registry controls, and investor reporting templates. The same metrics used here can then be applied directly: retrieval recall, complete citation accuracy, answer support, extractive faithfulness, and hallucination-risk after thresholding. The accompanying DOCX, dataset, scripts, metrics, and figures provide a complete replication package for the present findings and a template for that next benchmark.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. Yih, T. Rocktaschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. NeurIPS, 2020.

[2] Siming Zhao, Hailin Zhou, and Daniel Martinez, "LLM-Assisted Causal Attribution of Service Performance Upgrades on Churn and Tenure: Full Evaluation on the IBM Telco Customer Churn Dataset", JACS, vol. 3, no. 2, pp. 18–34, Feb. 2023, doi: 10.69987/JACS.2023.30202.

[3] 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.

[4] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," Communications of the ACM, vol. 18, no. 11, pp. 613-620, 1975.

[5] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," Journal of the American Society for Information Science, vol. 41, no. 6, pp. 391-407, 1990.

[6] R. Nogueira and K. Cho, "Passage Re-ranking with BERT," arXiv:1901.04085, 2019.

[7] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. EMNLP-IJCNLP, 2019.

[8] A. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, 2017.

[9] Yunhe Li, "Execution-Feedback and Retrieval-Augmented Generation for Conversational Text-to-SQL: From One-Shot Questions to Clarification-Driven Executable Dialogs", JACS, vol. 3, no. 2, pp. 1–17, Feb. 2023, doi: 10.69987/JACS.2023.30201.

[10] T. B. Brown et al., "Language Models are Few-Shot Learners," in Proc. NeurIPS, 2020.

[11] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A Large-scale Dataset for Fact Extraction and Verification," in Proc. NAACL-HLT, 2018.

[12] Yuanzheng Chen, Yitian Zhang, David Chau, and Matt Sherman, "Credit Card Default Risk Tiering with Probability Calibration and Uncertainty-Driven Rejection: A Reproducible Study on the UCI Credit Card Clients Dataset", JACS, vol. 3, no. 4, pp. 31–47, Apr. 2023, doi: 10.69987/JACS.2023.30403.

[13] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to Answer Open-Domain Questions," in Proc. ACL, 2017.

[14] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in Proc. ACL Workshop on Text Summarization Branches Out, 2004.

[15] T. Loughran and B. McDonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," Journal of Finance, vol. 66, no. 1, pp. 35-65, 2011.

[16] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," arXiv:1908.10063, 2019.

[17] U.S. Securities and Exchange Commission, "Commission Guidance Regarding Management's Discussion and Analysis of Financial Condition and Results of Operations," Release Nos. 33-8350 and 34-48960, 2003.

[18] U.S. Securities and Exchange Commission, "Modernization of Regulation S-K Items 101, 103, and 105," Release Nos. 33-10825 and 34-89670, 2020.

[19] U.S. Securities and Exchange Commission, "Management's Discussion and Analysis, Selected Financial Data, and Supplementary Financial Information," Release Nos. 33-10890 and 34-90459, 2020.

[20] U.S. Securities and Exchange Commission, "Asset-Backed Securities Disclosure and Registration," Release Nos. 33-9638 and 34-72982, 2014.

[21] Financial Accounting Standards Board, "Accounting Standards Update No. 2016-13: Financial Instruments-Credit Losses (Topic 326)," 2016.

[22] Financial Accounting Standards Board, "Accounting Standards Codification Topic 860: Transfers and Servicing," 2021.
[23] IFRS Foundation, "IFRS 9 Financial Instruments," 2014.
[24] World Economic Forum, "The Future of Financial Infrastructure: An Ambitious Look at How Blockchain Can Reshape Financial Services," 2016.
[25] Xinzhuo Sun, Yifei Lu, and Jing Chen, “Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting”, JACS, vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.