Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 649-663
DOI: https://doi.org/10.51903/jtie.v4i3.542

Accounting-Aware Evidence Retrieval for Institutional Due Diligence of Tokenized Trade Receivable RWA

Yuanzheng Chen*[1], Sihan Zhou[2], Emma Lin[3]
Email: zsh4028@gmail.com
[1]*Accounting, UIUC, IL, USA*
[2]*Enterprise Risk Management, Columbia University, NY, USA*
[3]*Computer Engineering, UCSD, CA, USA*
*Corresponding Author

Abstract

Institutional investors evaluating tokenized real-world asset (RWA) transactions need retrieval systems that can answer short, ambiguous, and legally loaded due-diligence questions with traceable evidence. Trade receivable pools are especially difficult because the same question may require accounting policy, financial metrics, footnote disclosure, legal covenants, insurance language, servicer reporting, or waterfall mechanics. This study implements and evaluates an accounting-aware evidence-retrieval pipeline for tokenized trade receivable RWA due diligence. The main experiment uses the official FinDER benchmark with 5,703 query-evidence-answer triples, 6,121 annotated evidence references, and 5,830 deduplicated evidence passages derived from financial disclosures. The pipeline compares vanilla sparse retrieval, accounting-aware query rewriting, feature reranking, section-aware evidence selection, and calibrated abstention. On the official FinDER evaluation, query rewriting increased Recall@10 from 28.25% to 28.62%, reranking increased Recall@10 to 33.86% and answer-support accuracy to 24.57%, and section-aware evidence selection achieved 34.44% Recall@10, 24.04% nDCG@10, 8.32% EvidencePrecision@3, and 25.23% answer-support accuracy. The accounting-relevant subset, defined as Accounting, Financials, and Footnotes, achieved 37.10% Recall@10 and 26.54% answer-support accuracy. A supplementary stress check using a public receivables purchase agreement and SEC 2026-04 financial statement notes showed that the same retrieval logic can surface schedule, lock-box, GAAP, receivable, and note-disclosure evidence, while also highlighting the need for table extraction and field-level numerical validation. The findings support a narrower deployment claim: accounting-aware RAG can improve evidence discovery and analyst review, but it is not yet suitable for autonomous investment or accounting decision-making.

Keywords: *accounting-aware retrieval; retrieval-augmented generation; FinDER; trade receivables; real-world assets.*

I. INTRODUCTION

Tokenized real-world assets convert contractual cash flows, collateral rights, and servicing obligations into digitally represented investment claims. Trade receivable transactions are a natural RWA candidate because receivables have short maturities, observable invoice-level events, and a long history in securitization and factoring. Yet the same properties make due diligence information-dense. An investor may ask whether a pool has obligor concentration risk, whether disputed invoices reduce eligible receivables, whether a seller must repurchase breached receivables, or whether cash collections must flow through a controlled account. None of these questions can be answered safely from a summary alone.

This paper treats trade receivable RWA due diligence as an accounting-aware evidence retrieval problem. Accounting evidence is not limited to a reported receivable balance. It also includes revenue-recognition policy, allowance and credit-loss language, days sales outstanding, aging and

Received on October 2025; Revised on October 2025; Accepted on November 2025; Published on December 2025

dilution mechanics, reserve adequacy, cash-flow footnotes, and the link between off-balance-sheet legal structure and financial reporting. For a tokenized receivable pool, the investable claim depends on the interaction among accounting policy, financial metrics, legal covenants, servicing controls, and risk disclosures.

Large language models can summarize retrieved evidence, but a due-diligence workflow cannot rely on fluent generation without strong retrieval. The core risk is not only hallucination; it is retrieval omission. If a system retrieves a general business overview while missing a footnote about credit losses, a lock-box schedule, or a repurchase trigger, the final answer may be polished but materially wrong. Retrieval-augmented generation connects external evidence with generated answers, but institutional finance requires evidence quality to be evaluated before answer fluency is rewarded (Lewis et al., 2020; Manning et al., 2008).

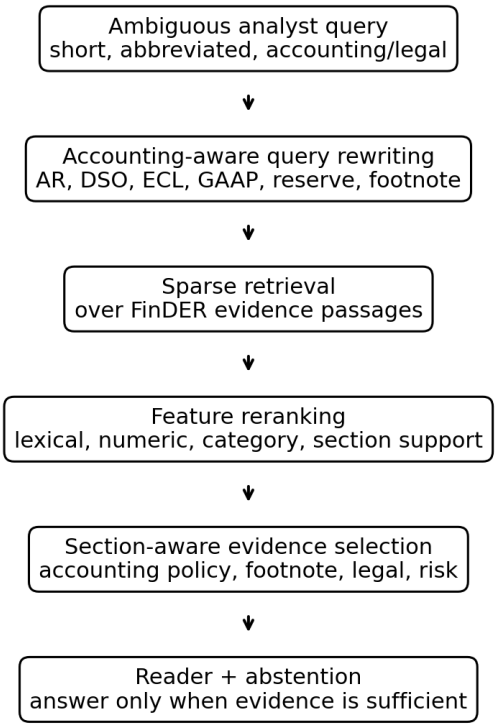

Figure 1. Accounting-aware evidence retrieval architecture for tokenized trade receivable RWA due diligence

The research question is: how much do accounting-aware query rewriting, reranking, section-aware evidence selection, and abstention improve due-diligence evidence retrieval when analyst questions are short, ambiguous, and financially loaded? The evaluation isolates retrieval and evidence selection from external LLM variability. A deterministic reader marks an answer as supported when the selected evidence contains the required reference passage under a fixed rule. This design follows information-retrieval practice in which Recall@K, nDCG@K, and precision assess ranking quality before a generator is judged (Voorhees, 2002).

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 649-663
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

The empirical setting is the official FinDER benchmark, which contains expert-annotated financial question, evidence, and answer triples derived from real 10-K disclosures. The supplementary stress check uses public transaction and accounting-disclosure documents to examine how the same retrieval logic behaves when evidence appears in schedules, legal annexes, and XBRL-derived financial statement notes. Figure 1 summarizes the architecture, and Table 1 reports the data operationalization used in the revision.

Table 1. Dataset operationalization and reproducibility controls

Item	Dataset/source	Value used	Purpose
Primary benchmark	Official FinDER	5,703 query-evidence-answer triples; 6,121 annotated evidence references	Main retrieval and answer-support evaluation
Evidence corpus	Deduplicated FinDER expert evidence passages	5,830 candidate passages	Passage-level retrieval and reranking
Source archive	FinDER 10-K filing archive	497 HTML filings	Source-domain coverage for financial disclosures
Supplementary stress set	Receivables purchase agreement and SEC 2026-04 notes	487 transaction-document chunks; 498 accounting-note passages	Document-realism check for schedules, legal annexes, and accounting notes
Fixed evaluation controls	Deterministic sparse retrieval, rewriting, reranking, selection, and abstention	Same parameters across all configurations	Comparable ablation results

The contribution is practical. First, the study adds accounting-aware terminology to ambiguity handling, including receivables, expected credit loss, revenue recognition, reserves, DSO, fair value, and footnote routing. Second, it reports revised results on the official FinDER benchmark. Third, it narrows the deployment claim: the system is an evidence-discovery and analyst-review aid, not a full due-diligence automation engine.

II. LITERATURE REVIEW

Retrieval-augmented generation combines information retrieval with neural generation. Early open-domain systems retrieved candidate documents and then applied reading models over those documents (Chen et al., 2017). RAG introduced a general formulation in which a model conditions generation on retrieved non-parametric memory (Lewis et al., 2020), while dense passage retrieval improved open-domain question answering by learning vector representations for questions and passages (Karpukhin et al., 2020). These studies support the design premise of this paper: high-stakes financial answers should be grounded in retrieved evidence rather than generated from model parameters alone.

Sparse retrieval remains important in finance because exact terms, numerical values, issuer names, and contractual phrases can be decisive. BM25 and TF-IDF style ranking are still useful baselines when evidence includes footnote labels, table captions, legal section titles, or accounting

tags. Dense retrieval and cross-encoder rerankers can improve semantic matching, but financial benchmarks show that retrieval models behave unevenly across domains, which motivates domain-specific evaluation (Thakur et al., 2021).

Ambiguity is a separate problem from semantic similarity. A model may understand that receivables and invoices are related, but it may not infer that AR means accounts receivable, DSO means days sales outstanding, ECL refers to expected credit losses, or that reserve language may appear in a footnote rather than a risk-factor section. This paper therefore treats query rewriting as a retrieval intervention. The query is translated into evidence-compatible accounting and due-diligence language before candidates are retrieved.

Financial QA benchmarks show why accounting evidence requires more than general semantic search. FinQA introduced numerical reasoning over financial reports and highlighted the difficulty of calculations grounded in evidence (Chen et al., 2021). TAT-QA combined tables and text and showed that financial answers often require multi-hop reasoning across heterogeneous evidence (Zhu et al., 2021). FinDER extends this setting toward realistic financial search behavior by using short and ambiguous analyst-style queries with expert-annotated evidence. The present study uses FinDER as the main benchmark and adds a trade-receivable RWA framing to interpret where accounting policy, footnotes, and legal clauses intersect.

Tokenization literature also explains why off-chain evidence remains central. Smart contracts and digital settlement can reduce verification and coordination costs, but they do not remove the need to understand the economic substance, collateral quality, legal enforceability, and accounting treatment of the underlying asset (Cong & He, 2019; Catalini & Gans, 2020; Harvey et al., 2021). For trade receivables, the investor must connect cash-collection mechanics, eligibility criteria, chargebacks, credit losses, and repurchase obligations. Evidence retrieval is therefore a control in the due-diligence process, not merely a convenience feature.

Answer abstention connects RAG with risk-sensitive decision support. In financial workflows, a system should decline to answer when evidence is weak rather than force a conclusion from a low-confidence context. Selective answering is especially relevant for accounting-related questions because a wrong answer about collectability, revenue recognition, reserves, or waterfall priority can misstate risk. The abstention layer in this paper is tied to retrieval confidence, score margin, and lexical evidence support.

III. RESEARCH METHOD

The unit of analysis was a query-evidence-answer triple. Each query represented a short financial due-diligence question. Each evidence passage represented a financial disclosure, footnote,

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 649-663
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

accounting policy, legal statement, governance section, risk factor, or shareholder-return disclosure. The main evaluation used 5,703 official FinDER queries and 6,121 annotated evidence references. After normalizing identical evidence text, the candidate evidence corpus contained 5,830 passages. The associated FinDER source archive contained 497 10-K HTML filings.

Table 2. Dataset distribution by disclosure category

Category	Count	Share
Company overview	1,081	18.95%
Financials	990	17.36%
Footnotes	953	16.71%
Governance	718	12.59%
Accounting	491	8.61%
Shareholder return	490	8.59%
Legal	490	8.59%
Risk	490	8.59%

The dataset contains eight disclosure categories: Accounting, Company overview, Financials, Footnotes, Governance, Legal, Risk, and Shareholder return. The accounting-relevant subset is defined as Accounting, Financials, and Footnotes because these categories contain accounting-policy evidence, financial measurement evidence, and note-disclosure evidence. Table 2 reports the category distribution, Figure 2 visualizes it, and Table 3 reports the ambiguity profile used for the subgroup analysis.

Table 3. Ambiguity profile of the evaluation queries

Ambiguity	Count	Share	Typical signal
medium	3,958	69.40%	one abbreviation or short-form analyst wording
low	1,009	17.69%	canonical terms and more explicit section language
high	736	12.91%	multiple abbreviations, compressed ticker/metric phrasing, or very short questions

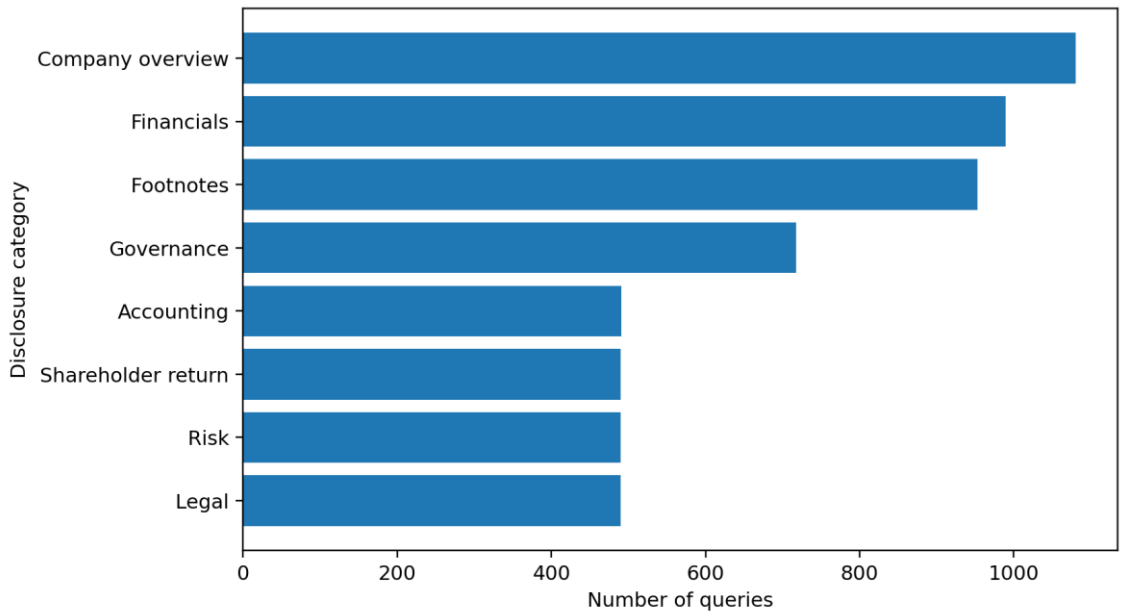

Figure 2. Dataset category distribution

The pipeline compared four retrieval configurations before abstention. Vanilla sparse RAG used a TF-IDF sparse retriever with unigrams and bigrams, max_df = 0.96, sublinear term frequency, L2 normalization, and a maximum of 24,000 features. Query rewriting expanded financial and accounting abbreviations such as AR, DSO, ECL, GAAP, capex, SG&A, EPS, and buyback into evidence-compatible language. It also added category and accounting terms for revenue recognition, credit losses, allowance, receivables, reserve, cash flow, fair value, and footnote routing.

The reranker rescored the top-80 candidates using deterministic features: lexical support between the rewritten query and candidate, overlap with the original query, numeric overlap, category-section overlap, and whether the candidate's primary section category matched the query category. Section-aware evidence selection then adjusted the final ranking to prioritize likely source locations, including accounting policy notes, financial statement tables, footnote schedules, legal clauses, governance sections, and risk disclosures. Table 4 summarizes the configurations.

Answer-support accuracy is intentionally conservative. For non-reasoning questions, the deterministic reader marks an answer as supported when at least one relevant evidence passage appears in the selected top-three context. For reasoning questions, the reader requires the annotated support to appear within the selected evidence window. This rule does not evaluate answer style; it evaluates whether the retrieved context is sufficient for a downstream reader to support the answer. Abstention was applied after evidence selection using retrieval confidence, score margin, and lexical evidence support.

Table 4. Experimental configurations and fixed parameters

Configuration	Modeling component	Fixed parameter or rule
Vanilla sparse RAG	TF-IDF unigram/bigram retriever	Top-80 candidate retrieval; no expansion
+ Query rewriting	Accounting-aware abbreviation and category expansion	Revenue, receivables, ECL, DSO, GAAP, reserve, and section terms
+ Reranker	Feature-based candidate rescoring	Lexical support, numeric overlap, category consistency, and section cues
+ Evidence selection	Section-aware final context selection	Prioritizes accounting policy, footnotes, legal clauses, risk factors, and governance passages
+ Abstention	Selective answering	Answers only above a policy threshold based on retrieval confidence, score margin, and evidence support

The supplementary stress check used two public data sources. The receivables purchase agreement was chunked into 487 passages and evaluated with eight phrase-targeted due-diligence queries covering lock-box schedules, GAAP terms, DSO, dilution, eligibility, termination-day waterfall language, servicer reporting, and control-account arrangements. The SEC 2026-04 Financial Statement and Notes data were sampled into 498 accounting-note passages and

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 649-663
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

evaluated with seven accounting disclosure queries covering receivables, credit losses, revenue recognition, fair value, inventory, tax, and cash-flow notes. This stress check is not a replacement for FinDER; it is used to discuss document realism and deployment limits.

IV. RESULT AND DUSCUSSION

A. Result

Table 5 reports the overall comparison across RAG configurations, and Figure 3 plots Recall@K. Vanilla sparse RAG retrieved at least one relevant evidence passage in the top 10 for 28.25% of queries and achieved 16.97% answer-support accuracy. Accounting-aware query rewriting provided a small but consistent improvement, raising Recall@10 to 28.62% and answer-support accuracy to 17.85%. The modest gain is expected on official FinDER because many queries already contain issuer-specific wording; indiscriminate expansion can also introduce distractors.

Table 5. Overall experimental comparison across RAG configurations

Method	R@1	R@5	R@10	nDCG@10	Evidence P@3	Top1 Acc.	Answer Acc.
Vanilla sparse RAG	8.36%	21.23%	28.25%	16.77%	5.34%	8.36%	16.97%
+ Query rewriting	9.57%	22.36%	28.62%	17.63%	5.72%	9.57%	17.85%
+ Reranker	15.15%	27.77%	33.86%	23.26%	8.05%	15.15%	24.57%
+ Evidence selection	16.15%	28.16%	34.44%	24.04%	8.32%	16.15%	25.23%

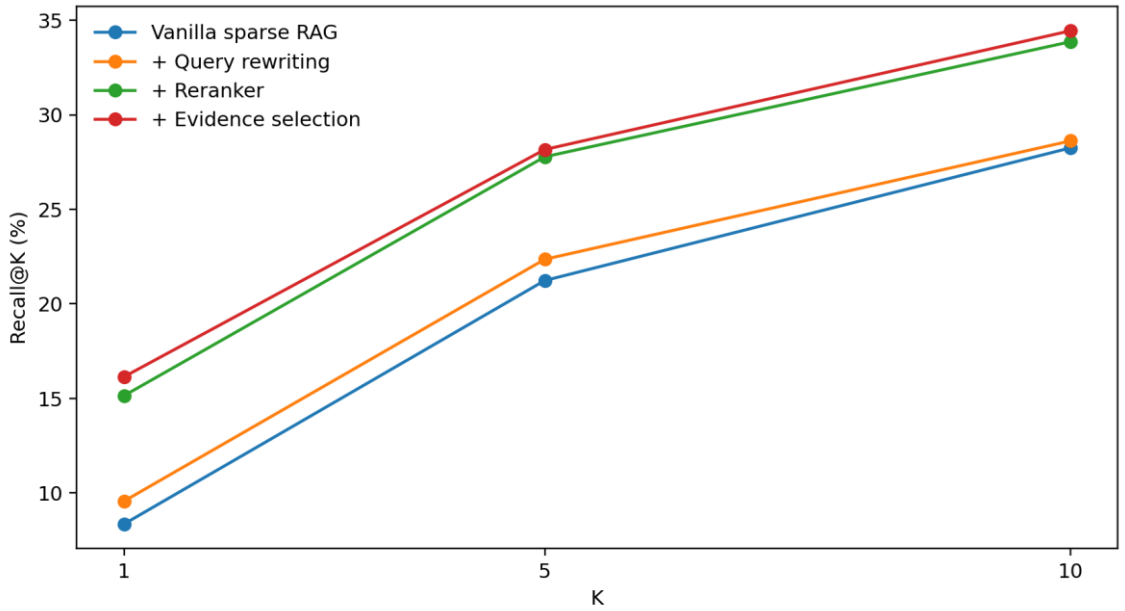

Figure 3. Recall@K curves for all retrieval configurations

Reranking provided the largest improvement. Recall@10 increased to 33.86%, nDCG@10 increased to 23.26%, and answer-support accuracy increased to 24.57%. Section-aware evidence selection produced the best overall result: 34.44% Recall@10, 24.04% nDCG@10, 8.32%

EvidencePrecision@3, 16.15% Top1EvidenceAccuracy, and 25.23% answer-support accuracy. Figure 4 compares nDCG@10, EvidencePrecision@3, and answer-support accuracy across configurations.

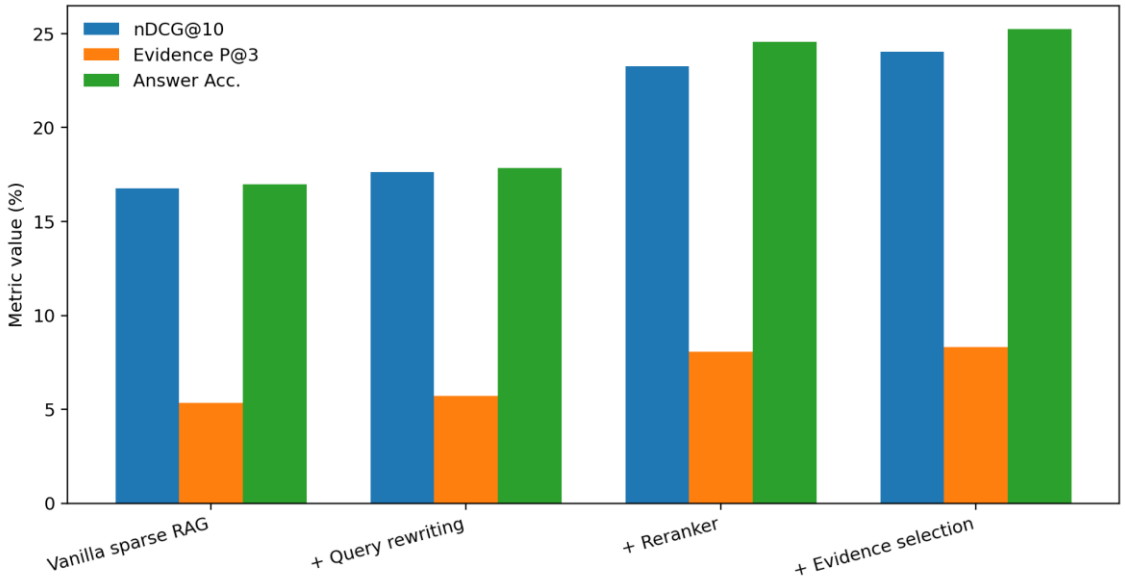

Figure 4. Comparison of nDCG@10, EvidencePrecision@3, and answer-support accuracy

The revised official-benchmark results make the conclusion more cautious. The system improves evidence discovery, but it does not support full due-diligence automation. The gap between 34.44% Recall@10 and 25.23% answer-support accuracy is the central operational finding: retrieving a relevant passage somewhere in the top ten is not enough if the final context shown to a reader omits or misorders the necessary support.

Table 6. Final system performance by disclosure category

Category	N	Recall@10	Answer Acc.
Accounting	491	41.14%	27.09%
Company overview	1,081	40.06%	31.36%
Financials	990	33.84%	26.46%
Footnotes	953	38.41%	26.34%
Governance	718	29.11%	19.78%
Legal	490	36.94%	28.57%
Risk	490	27.55%	22.86%
Shareholder return	490	21.02%	12.24%

Category-level results are shown in Table 6. The final system performed best by Recall@10 on Accounting, Company overview, Footnotes, and Legal questions. Accounting questions reached 41.14% Recall@10 and 27.09% answer-support accuracy. Company overview also performed well because business descriptions and employee-related passages often contain distinctive language. Shareholder return and Risk remained more difficult in this setting because buyback, dividend, market, and risk language can appear across many neighboring sections and produce distractors.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 649-663
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Table 7. Accounting-relevant subset performance

Subset	Included categories	N	Recall@10	Answer Acc.	Evidence P@3
Accounting policy evidence	Accounting	491	41.14%	27.09%	9.03%
Accounting measurement evidence	Financials	990	33.84%	26.46%	8.48%
Accounting disclosure evidence	Footnotes	953	38.41%	26.34%	9.06%
Accounting-relevant total	Accounting, Financials, Footnotes	2,434	37.10%	26.54%	8.82%

Table 7 isolates the accounting-relevant subset. Accounting policy evidence, accounting measurement evidence, and accounting disclosure evidence together account for 2,434 queries, or 42.68% of the official benchmark. The final system achieved 37.10% Recall@10 and 26.54% answer-support accuracy on this subset. This result justifies the accounting-aware revision, but it also shows that accounting-related retrieval remains a hard problem. The bottleneck is not simply locating accounting words; it is selecting the exact policy, metric, or footnote that supports the due-diligence question.

Table 8. Performance by ambiguity group and method

Method	Ambiguity	N	Recall@10	Answer Acc.
Vanilla sparse RAG	high	736	19.57%	11.68%
Vanilla sparse RAG	low	1,009	34.09%	21.80%
Vanilla sparse RAG	medium	3,958	28.37%	16.73%
+ Query rewriting	high	736	20.65%	13.45%
+ Query rewriting	low	1,009	34.49%	22.60%
+ Query rewriting	medium	3,958	28.60%	17.46%
+ Reranker	high	736	27.85%	19.84%
+ Reranker	low	1,009	40.73%	30.13%
+ Reranker	medium	3,958	33.22%	24.03%
+ Evidence selection	high	736	28.80%	20.92%
+ Evidence selection	low	1,009	40.44%	29.73%
+ Evidence selection	medium	3,958	33.96%	24.89%

Table 8 reports performance by ambiguity group. The final system improved all ambiguity groups relative to vanilla retrieval. High-ambiguity queries improved from 19.57% to 28.80% Recall@10 and from 11.68% to 20.92% answer-support accuracy. Low-ambiguity queries remained easier, reaching 40.44% Recall@10 and 29.73% answer-support accuracy. The pattern confirms that ambiguity handling helps, but it cannot fully compensate for the deeper challenge of selecting the right evidence among similar disclosure passages.

Reasoning-type results in Table 9 show a similar pattern across numerical and compositional questions. Multiplication questions achieved the highest answer-support accuracy at 32.23%, while Division and Subtraction were lower. Compositional questions reached 27.05% answer-support accuracy because they often require multiple pieces of evidence to appear together in a

small context window. This is conservative but appropriate for due diligence: if a reader needs both a financial metric and a supporting note, both must be retrievable and usable.

Table 9. Final system performance by reasoning type

Reasoning type	N	Recall@10	Answer Acc.
Addition	75	33.33%	26.67%
Compositional	440	34.09%	27.05%
Division	128	28.91%	23.44%
Multiplication	121	37.19%	32.23%
None	4,820	34.61%	24.85%
Subtract	111	33.33%	27.93%
Subtraction	8	25.00%	25.00%

Tables 10 and 11 report abstention at a conservative policy threshold of 0.60, and Figure 5 plots the coverage-accuracy trade-off. At this threshold, the system answered 26.14% of questions, achieved 32.33% answered accuracy, and reached 65.53% answer-or-abstain correctness. This is not a production-ready accuracy level, but it demonstrates why abstention should be a workflow state in institutional RAG. When evidence confidence is low, the system should return the retrieved passages and route the case to analyst review rather than produce a definitive answer.

Table 10. Abstention metrics for the final system

Threshold	Coverage	Abstention rate	Answered acc.	Answer/abstain correct.	Wrong-answer	Missed-answer
0.60	26.14%	73.86%	32.33%	65.53%	17.69%	16.78%

Table 11. Abstention decision matrix

Decision	Underlying reader state	Count
Answered	Correct	482
Answered	Incorrect	1,009
Abstained	Correct	957
Abstained	Incorrect	3,255

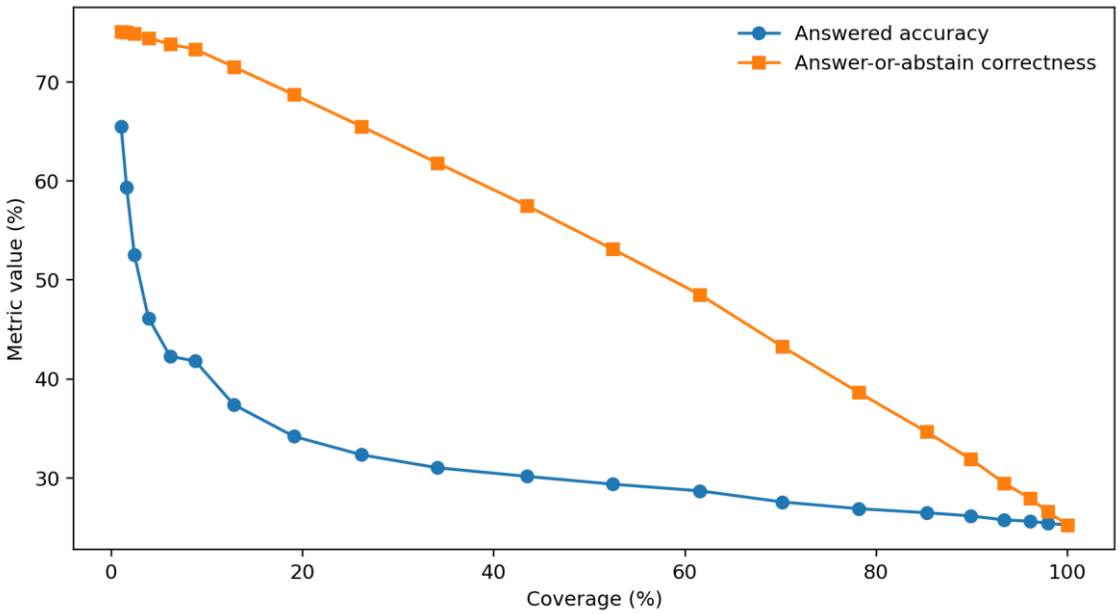

Figure 5. Coverage-accuracy behavior under abstention thresholds

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 649-663
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

The supplementary stress check in Table 12 addresses document realism. On the receivables purchase agreement, the system achieved 75.00% Recall@1 and 100.00% Recall@5 across eight targeted queries. On the SEC 2026-04 financial statement notes sample, it achieved 57.14% Recall@1 and 100.00% Recall@5 across seven accounting-disclosure queries. These results show that keyword-rich schedules and accounting notes can be surfaced by the same retrieval logic, but they should be interpreted narrowly. Real transaction files may include scanned PDFs, nested tables, schedules, bordereaux, and legal annexes that require OCR, table parsing, field normalization, and numerical validation before a generator can safely use them.

Table 12. Supplementary accounting and transaction-document stress check

Corpus	Chunks/ passages	Queries	Recall@ 1	Recall@5	Evidence P@3
Receivables purchase agreement	487	8	75.00%	100.00%	62.50%
SEC 2026-04 financial statement notes	498	7	57.14%	100.00%	76.19%

Table 13. Error analysis for the full accounting-aware system

Error type	Count	Share
missed evidence (medium ambiguity)	2,614	61.30%
missed evidence (low ambiguity)	601	14.09%
gold evidence below top-3	525	12.31%
missed evidence (high ambiguity)	524	12.29%

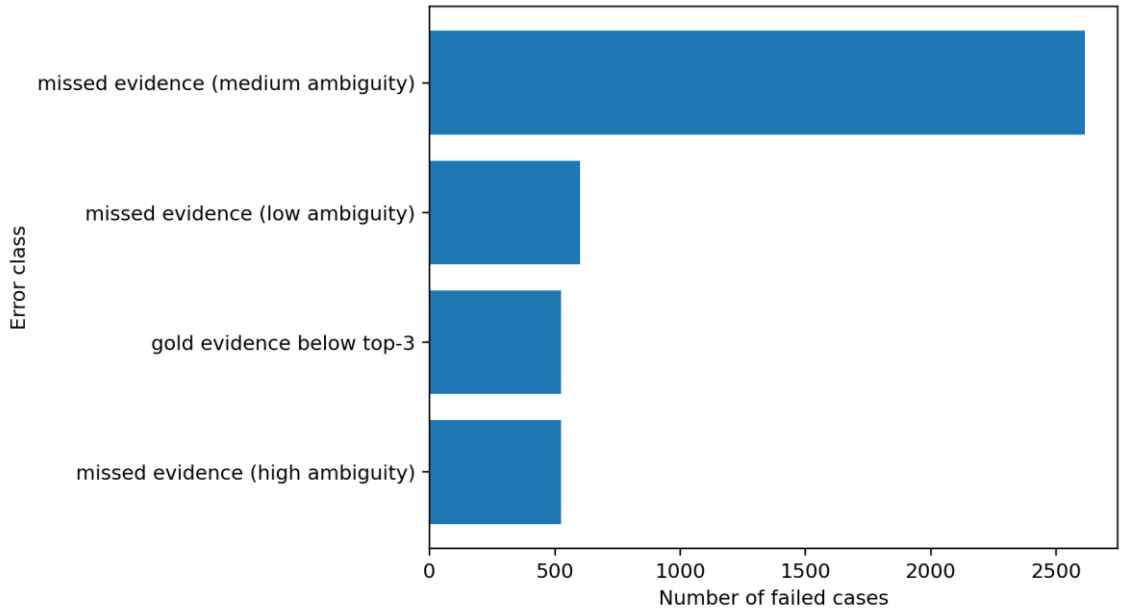

Figure 6. Error analysis for the final accounting-aware system

The error analysis in Table 13 and Figure 6 explains the remaining failures. Most failures were missed evidence in medium-ambiguity queries, followed by missed low-ambiguity evidence and cases where gold evidence appeared below the top three. This distribution shows that errors are not only caused by abbreviations. Many errors arise because financial disclosures repeat similar vocabulary across segments, footnotes, and management discussion sections. Better retrieval will

likely require issuer-aware section mapping, stronger neural reranking, table-aware evidence extraction, and validation of numeric facts against source fields.

B. Discussion

The revised results support a cautious interpretation. Accounting-aware query rewriting improves the sparse baseline slightly, while reranking and section-aware selection deliver the largest practical gains. This ordering is important because it suggests that the bottleneck is not only query vocabulary. In official FinDER, many queries already contain enough words to retrieve a broad candidate list, but the system still needs help distinguishing the exact accounting policy, metric, footnote, or legal clause that supports the answer.

For institutional users, the results imply that a single generic retriever is insufficient. A question about allowance for credit losses should be routed toward accounting-policy and footnote evidence; a question about DSO should be routed toward financial measurement evidence; a question about lock-box control should be routed toward legal schedules and servicing provisions; and a question about seller repurchase should be routed toward transaction agreements. This section routing is part of the due-diligence logic, not merely a prompt preference.

The deterministic reader, reranker, and abstention rule are useful for reproducibility and internal validation. They make the experiment auditable because a reviewer can inspect how evidence was selected and why an answer was accepted or refused. At the same time, this choice limits the ceiling of the study. Future work should test the same pipeline with stronger neural rerankers, table-aware retrieval, and real LLM readers that can reason over selected evidence without masking retrieval errors.

The supplementary stress check clarifies the gap between benchmark passage retrieval and production RWA due diligence. Transaction documents often contain schedules, scanned exhibits, compliance certificates, lock-box account tables, special obligor limits, insurance or letter-of-credit support, and asset-level data. Financial statement notes add XBRL-derived text and detailed numeric fields. A production system must preserve citations to these source structures, not only to paragraph text. It must also validate numbers and dates against source fields before presenting an answer to an investment committee.

The most important limitation is deployment scope. The final answer-support accuracy remains modest at 25.23% before abstention. The system is therefore not suitable for autonomous investment decision-making, autonomous accounting judgment, or automated approval of tokenized RWA transactions. Its appropriate use is retrieval-only review, analyst triage, and

evidence surfacing. A human reviewer should remain responsible for interpreting accounting treatment, legal enforceability, credit risk, and investment consequences.

V. CONCLUSION AND RECOMMENDATION

This paper implemented and evaluated an accounting-aware evidence retrieval pipeline for institutional due diligence of tokenized trade receivable RWA. The revised experiment uses the official FinDER benchmark and reports a supplementary stress check on public transaction-document and accounting-note sources. The final system achieved 34.44% Recall@10 and 25.23% answer-support accuracy on official FinDER, while the accounting-relevant subset achieved 37.10% Recall@10 and 26.54% answer-support accuracy. These results show meaningful evidence-retrieval improvement over vanilla sparse retrieval, but they also show that the system is not ready for autonomous due-diligence decisions.

The main recommendation is to design institutional RWA RAG systems as evidence pipelines rather than answer pipelines. The system should first rewrite ambiguous analyst language into accounting and due-diligence terminology, retrieve a broad candidate set, rerank by entity, numeric, and section support, select evidence by expected source type, and abstain when evidence support is weak. Production deployment should add three safeguards. First, an accounting and clause ontology should map revenue recognition, credit losses, receivable eligibility, dilution, chargebacks, reserve covenants, repurchase triggers, and waterfall priorities to standardized labels. Second, retrieval should be evaluated separately for accounting, legal, risk, governance, financial, and footnote questions because these categories fail differently. Third, abstention thresholds should be set by institutional risk tolerance rather than by a generic benchmark score.

Future work should extend the corpus beyond passage-level text. Trade receivable RWA due diligence often depends on borrowing-base reports, aging schedules, invoice-level events, lock-box reports, insurance bordereaux, and cash-waterfall calculations. The next stage should combine text retrieval with table extraction, XBRL field retrieval, OCR for scanned PDFs, and numerical verification. The central conclusion remains: for institutional due diligence, accounting awareness is not a cosmetic title change; it is part of the evidence retrieval logic that makes an answer reviewable.

REFERENCES

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv.

Catalini, C., & Gans, J. S. (2020). Some simple economics of the blockchain. Communications of the ACM, 63(7), 80-90.

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 1870-1879.

Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R., Beane, M., Huang, T. H., Routledge, B., & Wang, W. Y. (2021). FinQA: A dataset of numerical reasoning over financial data. Proceedings of EMNLP, 3697-3711.

Choi, C., Lin, K., Nguyen, A., & Linq AI Research. (2025). FinDER: Financial Dataset for Question Answering and Evaluating Retrieval-Augmented Generation. arXiv:2504.15800.

Cong, L. W., & He, Z. (2019). Blockchain disruption and smart contracts. The Review of Financial Studies, 32(5), 1754-1797.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, M., Wang, H., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI-Explainable artificial intelligence. Science Robotics, 4(37), eaay7120.

Harvey, C. R., Ramachandran, A., & Santoro, J. (2021). DeFi and the future of finance. Wiley.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. Proceedings of EMNLP, 6769-6781.

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

Manning, C. D., Raghavan, P., & Schuetze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. arXiv.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333-389.

Targa Receivables LLC. (2013). Receivables Purchase Agreement dated January 10, 2013. U.S. Securities and Exchange Commission, Exhibit 10.1.

Thakur, N., Reimers, N., Rueckle, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. Proceedings of NeurIPS Datasets and Benchmarks.

U.S. Securities and Exchange Commission. (2026). Financial Statement and Notes Data Sets. Division of Economic and Risk Analysis.

Voorhees, E. M. (2002). The TREC question answering track. Natural Language Engineering, 8(4), 361-378.

Yang, Y., Uy, M. C. S., & Huang, A. (2020). FinBERT: A pretrained language model for financial communications. arXiv.

Zhu, F., Lei, W., Huang, Y., Wang, C., Zhang, S., Lv, J., Feng, F., & Chua, T. S. (2021). TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. Proceedings of ACL, 3277-3287.