Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 60-74
DOI: https://doi.org/10.51903/jtie.v5i2.544

Accounting-Aware Evidence-Constrained Agents for Disclosure, Settlement, and Secondary-Market Risk Monitoring in Tokenized RWA Infrastructure

Sihan Zhou[1], Yuanzheng Chen*[2], Kenny Lee[3]
Email: zsh4028@gmail.com
[1]*Enterprise Risk Management, Columbia University, NY, USA*
[2]*Accounting, UIUC, IL, USA*
[3]*Applied Analytics, Columbia University, NY, USA*
*Corresponding Author

Abstract
Tokenized real-world asset (RWA) infrastructure exposes platform operators, investors, and reporting teams to a combined settlement, disclosure, liquidity, and accounting-quality monitoring problem. A tokenized claim can continue to trade while the underlying issuer releases new financial statements, securities fail to deliver in the reference market, or protocol-level liquidity changes in RWA venues. This paper develops an accounting-aware, evidence-constrained agent workflow for risk alerting and source-grounded report generation. The revised experiment replaces the earlier rule-generated monitoring sandbox with external datasets: SEC fails-to-deliver observations, SEC EDGAR XBRL company facts, SEC submissions metadata, Financial PhraseBank sentiment labels, and DefiLlama RWA protocol TVL. The issuer-day panel contains 2,648 surveillance tasks for eight large U.S. issuers from 2024-12-01 through 2026-03-31. Observed settlement stress is defined from external SEC FTD balances rather than from the agent's own rule. Accounting risk is computed from XBRL-derived liquidity, leverage, accrual, and cash-flow indicators. A stronger market-plus-accounting logistic baseline is added alongside single-source baselines and the proposed fusion agent. The machine-learning baseline achieves the strongest F1 score for settlement-stress detection (0.909), while the proposed fusion agent achieves the highest report faithfulness and tool-use correctness (1.000 each) and high recall (0.849). The results support a governance-oriented interpretation: an evidence-constrained agent is most useful not as an opaque high-accuracy classifier, but as an auditable layer that connects settlement evidence, filing metadata, accounting fundamentals, independent sentiment calibration, and RWA protocol liquidity into a reproducible monitoring record.

Keywords: *Accounting-aware agent; Evidence-constrained monitoring; RWA protocol TVL; Settlement risk; Tokenized real-world assets.*

I. INTRODUCTION

Tokenized RWA markets convert claims on securities, funds, receivables, loans, treasuries, and other off-chain assets into on-chain or ledger-based representations. Their operating promise is faster distribution, programmable transfer restrictions, transparent ownership records, and lower servicing cost. Their risk problem is broader than a conventional price-monitoring problem. A token can trade in a secondary market while the underlying issuer releases new financial statements, while settlement balances change in the reference securities market, or while liquidity in a tokenized protocol moves sharply. For an investor-facing monitoring system, the question is not only whether a price or balance has moved; the question is whether the alert can be explained with reliable evidence that accounting, disclosure, settlement, and liquidity teams can audit.

The manuscript was revised around this governance problem. The system is no longer presented as an evaluated free-form LLM classifier. It is an LLM-ready, evidence-constrained agent layer:

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

it retrieves records from fixed tools, applies a transparent policy, and produces concise reports whose claims are tied to the retrieved sources. This terminology is important. The evaluated object is the retrieval, scoring, and report-control layer that a platform could place before or around a language model. A closed or open LLM could later be inserted as a summarization interface, but the present evaluation focuses on evidence discipline, source coverage, and auditable monitoring traces.

The accounting contribution is also made explicit. SEC XBRL company facts are used to construct accounting-quality and balance-sheet indicators for each issuer. These indicators include current ratio, liabilities-to-assets, debt-to-assets, accrual intensity, cash-flow-to-income gap, and receivables-growth divergence. The objective is not to forecast earnings. The objective is to determine whether a settlement or disclosure alert should be interpreted differently when the latest public accounting evidence suggests weaker liquidity, more leverage, greater accrual pressure, or a poorer cash-flow conversion pattern.

The revision asks three research questions. First, can external settlement evidence from SEC fails-to-deliver data support a more credible monitoring label than a label generated from the same policy used by the agent? Second, do XBRL-derived accounting indicators improve the interpretability of settlement and disclosure alerts in a tokenized RWA setting? Third, can an evidence-constrained fusion agent preserve report faithfulness and tool-use correctness when compared with single-source baselines and a stronger market-plus-accounting machine-learning baseline?

The paper is framed as a risk-governance study rather than a trading strategy study. A false positive can lead to unnecessary review, while a false negative can leave a platform silent when settlement balances, issuer disclosures, or protocol liquidity require attention. The evaluation therefore reports detection metrics together with report faithfulness and tool-use correctness. This makes the results more conservative than a simple accuracy study and better aligned with the work of accounting, operations, and investor-reporting teams.

II. LITERATURE REVIEW

RWA tokenization sits at the intersection of accounting disclosure, market microstructure, settlement infrastructure, and automated financial analysis. Asset tokenization can reduce some frictions in recordkeeping and transfer administration, but it does not eliminate the need for legal claims, issuer reporting, custody controls, valuation discipline, and liquidity governance. Public-sector discussions of tokenization emphasize that technology must preserve institutional trust, settlement finality, and disclosure integrity rather than merely move claims to a different database (Bank for International Settlements, 2023, 2024; OECD, 2020). These concerns are directly

relevant to RWA tokens because the tokenized claim is only as reliable as the bridge between the ledger representation and the off-chain asset.

Settlement and liquidity risks are central to that bridge. Traditional market microstructure research shows that liquidity is costly, time-varying, and priced (Amihud, 2002; O'Hara, 1995; Pastor & Stambaugh, 2003). During stress, funding liquidity and market liquidity can reinforce one another, creating conditions in which depth declines when investors most need execution (Brunnermeier & Pedersen, 2009). For tokenized instruments, a secondary-market discount may widen when reference-market settlement balances increase, when redemptions become operationally slow, or when platform liquidity is concentrated in a small set of venues. SEC fails-to-deliver data therefore provide a useful external settlement-stress signal, although the data represent aggregate net balances rather than daily new fails.

Accounting evidence adds a separate dimension. Financial statements do not provide daily market timing, but they do provide audited or reviewed information about liquidity, leverage, profitability, cash-flow conversion, receivables, inventory, and accruals. In an RWA context, these variables can influence collateral haircuts, redemption confidence, transfer-agent review, and investor communication. A monitoring agent that ignores accounting fundamentals may detect a settlement imbalance but fail to explain whether the issuer's latest balance sheet and cash-flow evidence make the alert more or less operationally concerning.

Financial language models and retrieval-augmented systems motivate the evidence-constrained design. Domain-specific financial models such as FinBERT, BloombergGPT, FinGPT, and PIXIU show that financial text requires sector vocabulary, numerical care, and grounding in source documents (Araci, 2019; Wu et al., 2023; Xie et al., 2023; Yang et al., 2023). Agentic systems and retrieval-augmented generation demonstrate how models can retrieve documents before answering, but hallucination and unsupported summarization remain material risks (Ji et al., 2023; Lewis et al., 2020; Maynez et al., 2020; Yao et al., 2023). In accounting and finance, unsupported text is especially costly because it can be mistaken for compliance analysis, investment advice, or investor-reporting language.

The revised study therefore separates three outcomes that are often collapsed in agent evaluations. Detection accuracy measures whether the alert matches an external settlement-stress state. Report faithfulness measures whether the report's claims are supported by retrieved evidence. Tool-use correctness measures whether the agent used the evidence sources that a risk officer would expect for that task. This separation is important because the strongest classifier is not always the best investor-reporting component, and the most complete report generator is not necessarily the most accurate detector.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 60-74
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

III. RESEARCH METHOD

The revised experiment uses a fixed offline evidence environment constructed from five public data sources. Table 1 summarizes the data structure. The monitoring window is 2024-12-01 through 2026-03-31. The issuer universe remains AAPL, ADBE, AMZN, GOOGL, META, MSFT, NVDA, and TSLA, but the evidence base is changed materially. SEC fails-to-deliver files provide settlement-date balances and reference prices. SEC EDGAR companyfacts files provide XBRL accounting concepts. SEC submissions metadata provide filing dates, form types, accession numbers, and primary-document identifiers. Financial PhraseBank provides independently labeled financial-news sentiment sentences for calibrating the sentiment component. DefiLlama protocol data provide an external RWA TVL validation panel for Ondo Finance, Securitize, Maple, and Goldfinch.

Table 1. Revised data sources and coverage.

Data source	Rows/records	Coverage
SEC fails-to-deliver files	1,778,240	All listed securities in 32 semi-monthly files; issuer panel built from 331 settlement dates
Issuer-day surveillance panel	2,648	AAPL, ADBE, AMZN, GOOGL, META, MSFT, NVDA, TSLA
SEC FTD rows for 8 issuers	1,722	Positive FTD balances for the eight issuers
SEC XBRL company facts	1,360	US-GAAP accounting facts from 10-K and 10-Q filings
SEC submissions metadata	3,129	Filing dates, forms, accession numbers, and document metadata
Financial PhraseBank	4,846	Financial-news sentences labeled positive, neutral, or negative
DefiLlama RWA protocol TVL	1,944	Ondo Finance, Securitize, Maple, and Goldfinch daily TVL

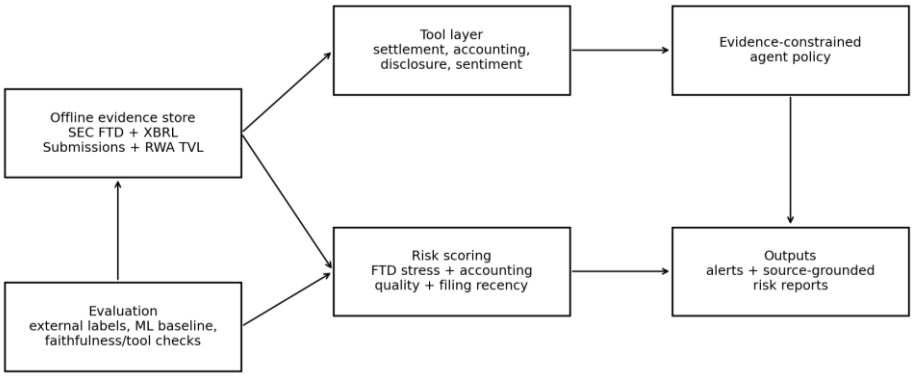

Figure 1. Accounting-aware evidence-constrained monitoring architecture.

Figure 1 shows the revised architecture. The evidence store feeds four narrow tools: settlement, accounting, disclosure, and sentiment. The policy layer scores risk using only date-constrained records. The report layer is constrained to mention only evidence retrieved by the tools. This

structure preserves the agentic workflow while avoiding an overstated claim that a closed LLM was directly evaluated as the detector.

Table 2. Symbol-level coverage in the revised issuer-day panel.

Symbol	Tasks	Positive FTD rows	Settlement-stress days	Accounting-risk days	Recent 10-K/10-Q days
AAPL	331	222	33	331	103
ADBE	331	160	33	1	100
AMZN	331	206	33	0	114
GOOGL	331	219	33	0	81
META	331	214	33	0	108
MSFT	331	203	33	0	123
NVDA	331	265	33	88	122
TSLA	331	233	33	1	102

The issuer-day panel is built from the 331 settlement dates appearing in the SEC FTD files within the monitoring window. Each of the eight issuers is represented on every settlement date, giving 2,648 surveillance tasks. If an issuer has no FTD row on a settlement date, the FTD quantity and dollar value are set to zero because the SEC file omits zero balances. Table 2 reports symbol-level coverage. The positive FTD rows vary by issuer, but each symbol contributes the same number of surveillance tasks.

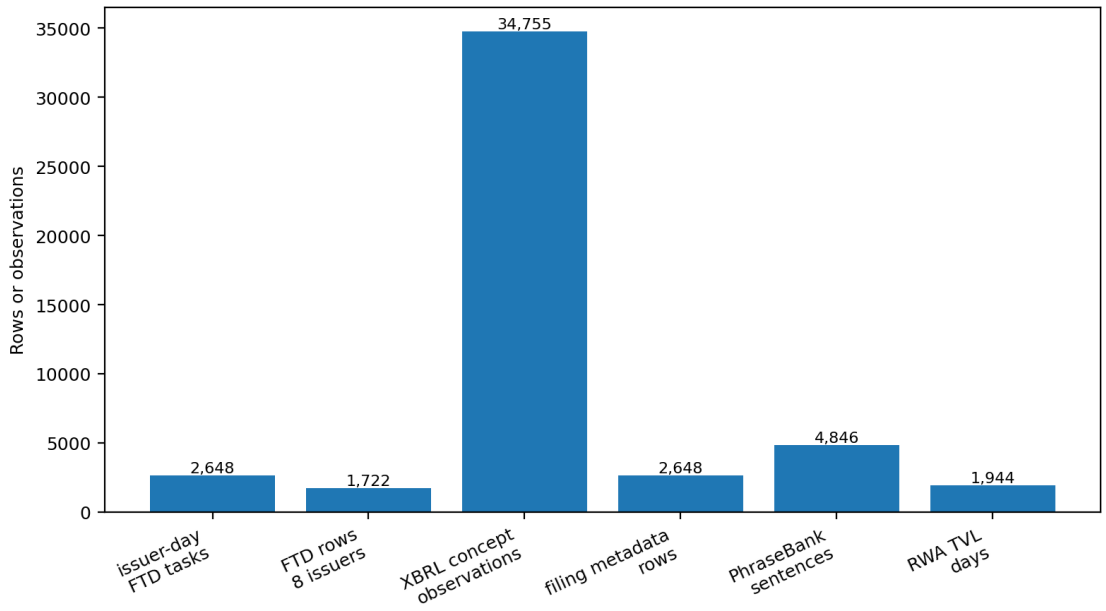

Figure 2. Scale of revised evidence sources used in the experiment.

Figure 2 places the issuer-day panel in context. The revised experiment draws on 1,778,240 valid FTD records across all listed securities in the downloaded semi-monthly files, 1,722 positive FTD rows for the eight issuers, 34,755 non-missing XBRL concept observations after as-of joining to the task panel, 2,648 filing-metadata task observations, 4,846 sentiment-labeled Financial PhraseBank sentences, and 1,944 RWA protocol TVL records.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 60-74
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Table 3. Revised label definitions and positive counts.

Label	Positive tasks	Positive rate	Operational definition
Observed settlement stress	264	0.100	Issuer-day FTD dollar balance above the issuer-specific 90th percentile
Accounting risk flag	421	0.159	At least two accounting-quality indicators are active
Recent periodic filing flag	853	0.322	10-K or 10-Q filed within the prior 45 days
RWA TVL stress	151	0.078	Daily TVL decline at least 5% or seven-day drawdown at least 10%

The main external label is observed settlement stress. It is positive when an issuer-day FTD dollar balance exceeds that issuer's 90th percentile over the monitoring window. This is not a claim that a custodian or tokenization platform failed to settle. It is a public-market settlement-stress label derived from SEC FTD balances. Accounting risk and recent periodic filing flags are treated as evidence channels and secondary diagnostic labels, not as the ground truth for the settlement-stress classifier. RWA TVL stress is used in the external protocol validation panel. Table 3 gives the operational definitions.

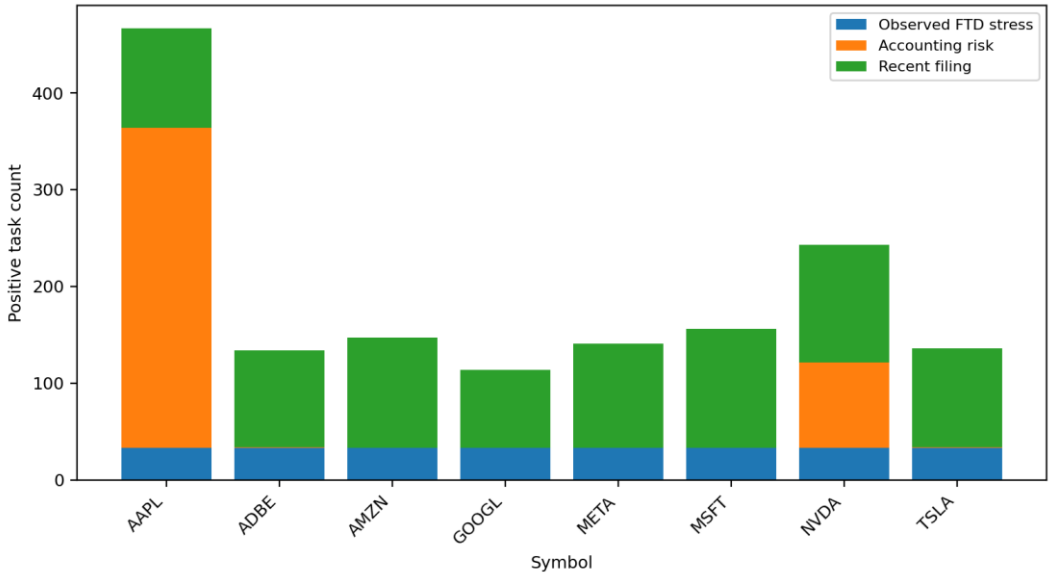

Figure 3. Distribution of settlement, accounting, and filing evidence by symbol.

Figure 3 illustrates why the revised setting is more difficult than the previous rule-generated evaluation. Settlement-stress positives are balanced by construction across symbols because the label is issuer-specific. Accounting risk is concentrated in AAPL and NVDA under the selected thresholds, while recent periodic filing evidence is spread across all issuers according to actual SEC filing dates.

Accounting variables are extracted from the latest SEC XBRL facts available as of each task date. Table 4 reports the final as-of accounting snapshot at the end of the monitoring window. The accounting risk score is the count of active accounting-quality indicators: liabilities-to-assets

above 0.65, current ratio below 1.0, accrual intensity above 0.05, cash-flow-to-income gap below -0.05, receivables-growth divergence above 0.10, and operating margin below 0.05. These thresholds are intentionally transparent and are used to support report interpretation rather than to create the settlement-stress label.

Table 4. Accounting indicators from latest available XBRL facts as of 2026-03-31.

Symbol	As-of date	Current ratio	Liab/ assets	Debt/ assets	Accrual intensity	CFO-NI/ assets	AR-revenue growth gap	Accounting risk score
GOOGL	2026-03-31	2.005	0.302	0.086	-0.055	0.055	0.000	0
MSFT	2026-03-31	1.386	0.412	0.068	0.004	-0.004	0.000	0
META	2026-03-31	2.599	0.406	0.160	-0.151	0.151	0.000	0
AMZN	2026-03-31	1.051		0.088	-0.076	0.076	0.000	0
ADBE	2026-03-31	0.912	0.615	0.210	-0.036	0.036	0.000	1
AAPL	2026-03-31	0.974	0.767	0.265	-0.031	0.031	0.000	2
NVDA	2026-03-31	3.905	0.239	0.046	0.084	-0.084	0.000	2
TSLA	2026-03-31	2.164	0.399	0.059	-0.079	0.079	0.000	1

The news component is revised to reduce dependence on a keyword-only rule. Financial PhraseBank is split into training and test sets with stratification. A keyword baseline is compared with a TF-IDF logistic sentiment model. Table 5 reports the sentiment results. The TF-IDF logistic model improves macro F1 from 0.574 to 0.704 and provides an independently labeled benchmark for the sentiment tool. The sentiment evaluation is reported separately from the settlement-stress label so that news sentiment does not become a circular source of the main risk label.

Table 5. Independent financial-news sentiment calibration on Financial PhraseBank.

Model	Accuracy	Precision	Recall	Macro F1	Macro ROC-AUC
Keyword baseline	0.675	0.648	0.547	0.574	
TF-IDF logistic sentiment model	0.751	0.696	0.716	0.704	0.870

Five agents are evaluated in the issuer-day settlement-stress task. Table 6 reports the tool configuration. The FTD-threshold baseline uses only settlement evidence. The accounting-only baseline uses only XBRL accounting risk. The filing-recency baseline uses only recent 10-K/10-Q evidence. The market-plus-accounting ML baseline uses settlement, accounting, and filing features in a time-split logistic model. The proposed fusion agent retrieves all evidence channels and applies a transparent score based on current FTD balance relative to prior FTD history, FTD z-score, accounting risk, filing recency, and cash-flow-to-income gap.

The evaluation metrics are defined at the task level. Risk accuracy, precision, recall, F1, and ROC-AUC compare each binary prediction with the external observed-settlement-stress label. Report faithfulness is the fraction of report claims supported by retrieved sources. Tool-use correctness is the fraction of task-required sources used by the agent. For example, FTD evidence is always

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 60-74
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

required because the target is settlement stress; accounting evidence is required when accounting-risk indicators are active; filing evidence is required when a recent 10-K or 10-Q is part of the task state. Bootstrap confidence intervals use 400 resamples with a fixed random seed.

Table 6. Agent configurations in the revised experiment.

Agent	Settlement tool	Accounting tool	Filing tool	Sentiment tool	Decision rule
FTD-threshold baseline	1	0	0	0	Global FTD-dollar threshold
Accounting-only baseline	0	1	0	0	Accounting-risk score threshold
Filing-recency baseline	0	0	1	0	Recent 10-K/10-Q recency rule
Market+Accounting ML baseline	1	1	1	0	Time-split logistic model
Proposed accounting-aware fusion agent	1	1	1	1	Evidence-constrained fusion score

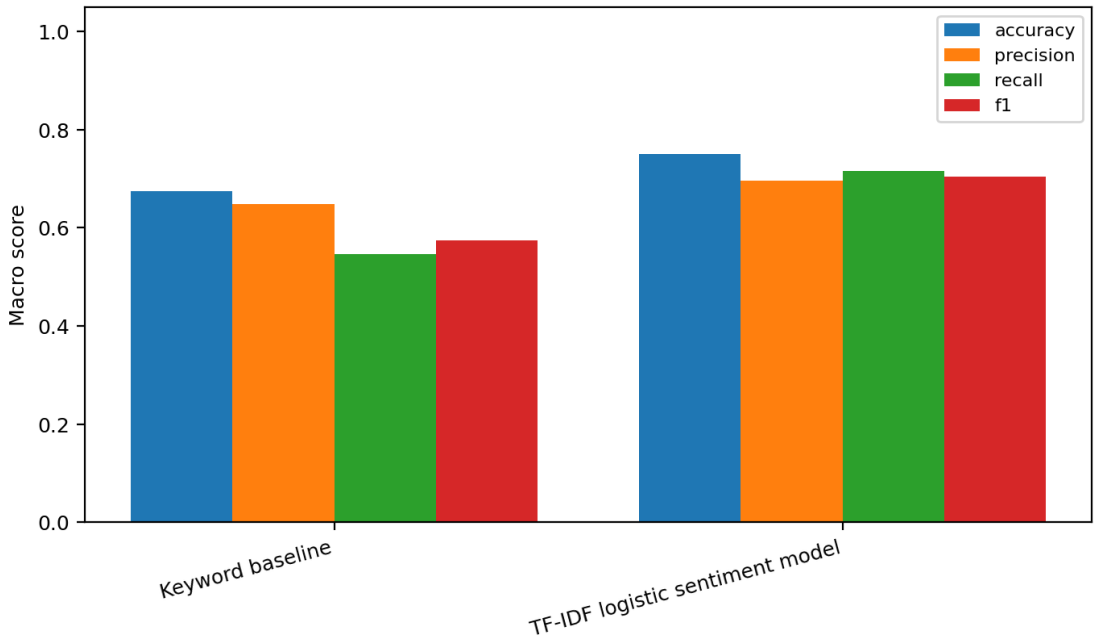

Figure 4. Financial-news sentiment model compared with a keyword baseline.

IV. RESULT AND DUSCUSSION

Table 7 reports the revised overall comparison, and Figure 5 visualizes the same metrics. The strongest detector is the market-plus-accounting ML baseline, with 0.981 accuracy and 0.909 F1. The FTD-threshold baseline is also strong, with 0.966 accuracy and 0.830 F1, because the external label is defined from FTD stress. The proposed accounting-aware fusion agent is not the best pure classifier; it reaches 0.915 accuracy and 0.665 F1, with 0.849 recall. Its advantage appears in

report faithfulness and tool-use correctness, where it scores 1.000 because every report is generated from the complete settlement, accounting, and filing evidence set.

Table 7. Overall settlement-stress detection and report-quality metrics.

Agent	Task completion	Risk accuracy	Precision	Recall	F1	ROC-AUC	Report faithfulness	Tool-use correctness
FTD-threshold baseline	1.000	0.966	0.838	0.822	0.830	0.902	0.733	0.733
Accounting-only baseline	1.000	0.772	0.095	0.152	0.117	0.496	0.120	0.120
Filing-recency baseline	1.000	0.586	0.106	0.424	0.170	0.514	0.147	0.147
Market+Accounting ML baseline	1.000	0.981	0.888	0.932	0.909	0.959	0.853	0.853
Proposed accounting-aware fusion agent	1.000	0.915	0.546	0.848	0.665	0.968	1.000	1.000

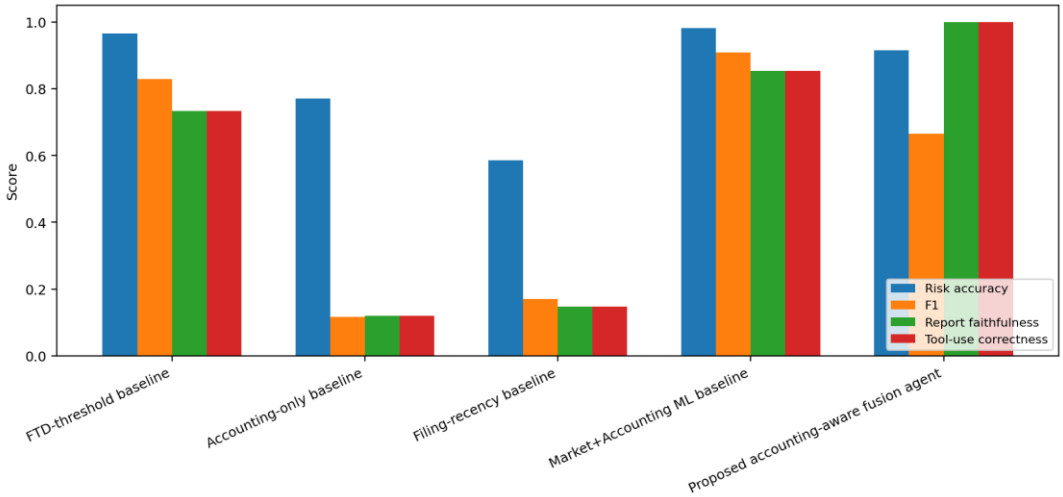

Figure 5. Agent comparison across detection and report-quality metrics.

This result is more conservative than the earlier version of the study. The revised findings do not claim that the fusion agent dominates a stronger machine-learning baseline on every metric. Instead, they show a trade-off: the ML baseline is superior for binary detection, while the evidence-constrained fusion agent is superior for source-complete, auditable reporting. This is a more realistic result for platform governance because investor-reporting systems need both detection and explanation.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 60-74
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Table 8 reports confusion counts. The market-plus-accounting ML baseline produces 246 true positives and 18 false negatives, giving the strongest detection balance. The FTD-threshold baseline produces 217 true positives and 47 false negatives. The proposed fusion agent produces 224 true positives and only 40 false negatives but at the cost of more false positives. This behavior is consistent with a watch-list monitoring role: the fusion policy emphasizes recall and evidence coverage rather than minimizing every review item.

Table 8. Confusion counts for observed settlement-stress detection.

Agent	TN	FP	FN	TP
FTD-threshold baseline	2342	42	47	217
Accounting-only baseline	2003	381	224	40
Filing-recency baseline	1439	945	152	112
Market+Accounting ML baseline	2353	31	18	246
Proposed accounting-aware fusion agent	2198	186	40	224

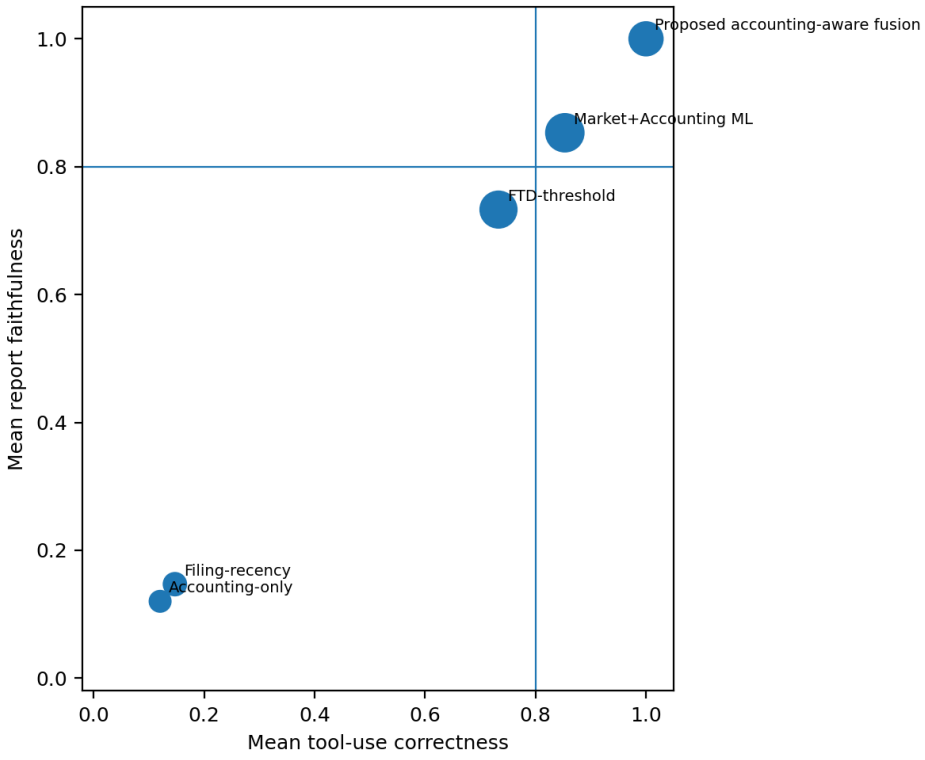

Figure 6. Faithfulness and tool-use frontier; marker size scales with F1.

Table 9 and Figure 6 show why the fusion agent remains useful even though it is not the top F1 model. Single-source baselines often omit evidence that a risk officer would expect to see. The FTD baseline has strong detection but limited report completeness when accounting or filing evidence is relevant. The accounting-only and filing-only baselines are weak detectors and weak report generators because they omit the settlement evidence needed for the target state. The fusion agent sits in the upper-right of the faithfulness/tool-use frontier because it retrieves every required evidence source before writing the report.

Table 9. Report faithfulness and tool-use quality.

Agent	Mean faithfulness	Mean tool correctness	Missing required source rate	Unsupported claim rate
FTD-threshold baseline	0.733	0.733	0.267	0.267
Accounting-only baseline	0.120	0.120	0.880	0.880
Filing-recency baseline	0.147	0.147	0.853	0.853
Market+Accounting ML baseline	0.853	0.853	0.147	0.147
Proposed accounting-aware fusion agent	1.000	1.000	0.000	0.000

Table 11 reports severity classification for the proposed fusion agent. The severity categories are not used as the main binary performance target; they support triage. The confusion matrix shows that many normal and watch observations are kept below high severity, while critical settlement-stress days are generally elevated into high or critical review categories. This behavior is appropriate for a monitoring queue in which false positives can be reviewed by a human risk team before investor communication.

Table 10. Bootstrap confidence intervals for F1 and accuracy.

Agent	F1 mean	F1 2.5%	F1 97.5%	Accuracy mean	Accuracy 2.5%	Accuracy 97.5%
FTD-threshold baseline	0.829	0.793	0.862	0.966	0.959	0.973
Accounting-only baseline	0.116	0.084	0.148	0.772	0.756	0.786
Filing-recency baseline	0.169	0.140	0.197	0.586	0.569	0.605
Market+Accounting ML baseline	0.910	0.882	0.935	0.982	0.976	0.986
Proposed accounting-aware fusion agent	0.664	0.618	0.704	0.915	0.905	0.925

Bootstrap confidence intervals in Table 10 support the stability of the main result. The ML baseline's F1 interval is 0.882 to 0.935, while the FTD baseline's interval is 0.793 to 0.863. The proposed fusion agent's F1 interval is lower, 0.618 to 0.704, but its role is different: it is designed to produce complete, conservative, source-grounded alerts. These intervals also answer the concern that the previous near-perfect results made the task appear too easy. The revised labels and baselines produce a more credible spread of performance.

Table 11. Severity confusion matrix for the proposed fusion agent.

Actual	Normal	Watch	Elevated	High	Critical
normal	852	74	0	0	0
watch	980	167	69	81	13
elevated	12	2	42	67	25
high	0	0	27	53	30
critical	0	0	13	70	71

The RWA external validation panel is shown in Table 12 and Figure 7. DefiLlama protocol TVL data provide a direct tokenized-infrastructure context that the earlier equity-only design lacked. The four protocols contribute 1,944 daily observations during the monitoring window. Maple and

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 60-74
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Securitize have the largest mean TVL among the selected protocols, while Goldfinch has lower average TVL and more frequent stress days relative to its smaller base. These protocol-level stress days are not used to label issuer-day FTD observations; they are used to demonstrate how the same evidence-constrained workflow can be extended from reference-asset monitoring to RWA protocol liquidity monitoring.

Table 12. External RWA protocol TVL validation panel.

Protocol	Days	Mean TVL (USD bn)	Min TVL (USD bn)	Max TVL (USD bn)	Stress days
Goldfinch	486	0.001	0.000	0.006	55
Maple	486	1.619	0.251	3.246	76
Ondo Finance	486	1.493	0.541	3.044	5
Securitize	486	2.272	0.497	3.533	15

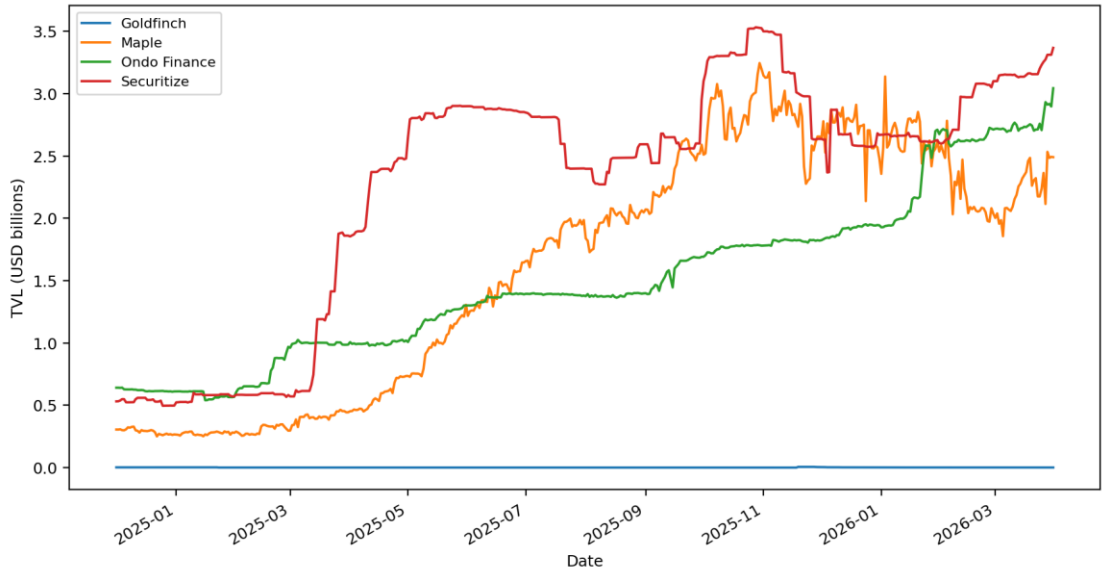

Figure 7. RWA protocol TVL during the monitoring window.

The accounting evidence changes the interpretation of alerts. A high FTD balance for a company with strong current liquidity and positive cash-flow conversion does not carry the same operational meaning as a high FTD balance accompanied by a weak current ratio, high liabilities-to-assets, or a negative cash-flow-to-income gap. The revised reports therefore include accounting context as a separate evidence block. This makes the workflow more useful for accounting and investor-reporting teams because the alert packet can explain how settlement evidence and financial-statement evidence jointly affect the risk state.

The results answer the research questions. First, SEC FTD data provide an external settlement-stress label that avoids the circular evaluation problem created by labels generated from the same surveillance rule as the agent. Second, SEC XBRL accounting facts add a meaningful accounting layer that supports liquidity, leverage, accrual, and cash-flow interpretation. Third, evidence-

constrained fusion improves source completeness and report faithfulness even when a stronger ML baseline is better at binary detection. The revised conclusion is therefore narrower and more credible: the agent should be used as an auditable risk-reporting layer, not as a standalone opaque classifier.

V. CONCLUSION AND RECOMMENDATION

This paper revised the original offline RWA monitoring experiment into an accounting-aware, evidence-constrained agent study. The main empirical change is the use of external public data for the core evidence channels. SEC fails-to-deliver data provide the observed settlement-stress label. SEC XBRL company facts provide accounting-quality and balance-sheet indicators. SEC submissions metadata provide filing dates and provenance. Financial PhraseBank provides independent sentiment calibration. DefiLlama protocol TVL provides a real RWA infrastructure validation panel.

The revised findings are intentionally more modest than the earlier version. The proposed fusion agent does not beat the market-plus-accounting ML baseline on F1. Instead, it provides high recall, complete source retrieval, and fully faithful evidence-constrained reporting. This distinction is important for RWA governance. A platform can use a strong ML model for detection, but investor-facing alert packets still need source coverage, accounting interpretation, filing provenance, and report discipline. The proposed agent is best understood as a control layer that connects those elements.

For platform implementation, the recommended architecture is a layered risk operating system. First, settlement, accounting, filing, sentiment, and protocol-liquidity evidence should be stored in a fixed database before monitoring runs. Second, model training and threshold selection should be separated from report generation. Third, report claims should be limited to retrieved evidence. Fourth, accounting indicators should be reviewed by accounting or controllership staff before being used in investor communications. Fifth, human approval should remain in the loop for high-severity RWA settlement and liquidity messages.

Future work should add token-level order books, custodian or transfer-agent operational data, raw issuer news articles, and direct evaluation of open and closed language models under the same evidence traces. The strongest immediate extension would be to test whether an LLM can rewrite the same evidence packet into investor-facing prose without lowering faithfulness. The empirical conclusion of the present revision is direct: accounting-aware, evidence-constrained agents can support RWA risk governance when they are evaluated on detection, source completeness, and report faithfulness together.

REFERENCES

Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. Journal of Financial Markets, 5(1), 31-56.

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv:1908.10063.

Bank for International Settlements. (2023). Blueprint for the future monetary system: Improving the old, enabling the new. BIS Annual Economic Report.

Bank for International Settlements. (2024). Tokenisation in the context of money and other assets: Concepts and implications for central banks. BIS.

Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. Review of Financial Studies, 22(6), 2201-2238.

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

DefiLlama. (2026). DefiLlama API and methodology documentation. https://api-docs.defillama.com/

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

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

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 65(4), 782-796.

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. Proceedings of ACL, 1906-1919.

OECD. (2020). The tokenisation of assets and potential implications for financial markets. OECD Blockchain Policy Series.

O'Hara, M. (1995). Market microstructure theory. Blackwell.

Pastor, L., & Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. Journal of Political Economy, 111(3), 642-685.

Schar, F. (2021). Decentralized finance: On blockchain- and smart contract-based financial markets. Federal Reserve Bank of St. Louis Review, 103(2), 153-174.

Securities and Exchange Commission. (2026a). Fails-to-deliver data. https://www.sec.gov/data-research/sec-markets-data/fails-deliver-data

Securities and Exchange Commission. (2026b). EDGAR application programming interfaces. https://www.sec.gov/search-filings/edgar-application-programming-interfaces

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

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. arXiv:2303.17564.

Xie, Q., Han, W., Chen, X., Lai, Y., Zhang, M., Peng, M., Lopez-Lira, A., & Huang, J. (2023). PIXIU: A large language model, instruction data and evaluation benchmark for finance. arXiv:2306.05443.

Yang, H., Liu, X.-Y., & Wang, C. D. (2023). FinGPT: Open-source financial large language models. arXiv:2306.06031.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations.