Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 119-141
DOI: https://doi.org/10.51903/jtie.v5i2.566

LLM-Inspired Ontology-Based Semantic Enrichment for FinTech M&A Intelligence in SEC Structured Disclosures

Dingyuan Zhang[1†], Guanzheng Zhao[2*†], Sisi Meng[1]
Email: gzhaolance14@gmail.com
[1]Business Analytics / Accounting, University of Rochester, NY, USA
[2]Financial Engineering, Stevens Institute of Technology, NJ, USA
*† Dingyuan Zhang and Guanzheng Zhao contributed equally. * Corresponding author: Guanzheng Zhao.*

Abstract

This study develops an ontology-based event-intelligence framework for FinTech merger-and-acquisition evidence in U.S. Securities and Exchange Commission (SEC) structured disclosures. Five quarterly SEC Financial Statement Data Sets from 2025Q1 through 2026Q1 contain 32,254 filings, 7,335 registrants, and 18,312,494 numerical XBRL facts; validation adds a 1,800-filing full-text Form 8-K sample and 192 FDIC events. A deterministic ontology maps XBRL tag names, labels, and documentation to acquisition, disposition, valuation, integration, risk, and payment concepts. The method is therefore LLM-inspired semantic enrichment rather than direct LLM extraction. Logistic regression, decision tree, and random forest classifiers are evaluated in three expanding forward-quarter tests with training-only preprocessing and threshold selection. A proximal protocol retains semantically close predictors, whereas a strict protocol excludes label-generating variables and deterministic descendants. Across the rolling tests, proximal logistic-regression M&A detection attains mean F1 = 0.981941, ROC-AUC = 0.999517, and average precision = 0.998315; strict performance falls to F1 = 0.735640, ROC-AUC = 0.930870, and average precision = 0.824107. FinTech M&A F1 declines from 0.921198 to 0.449503. Strict random forests yield F1 = 0.798129 for integration risk and 0.753987 for valuation signals. In independent full text, strict main-text-plus-exhibit F1 is 0.297482; among 24 automatically linked FDIC events in rolling test quarters, 9 are detected. Near-perfect scores therefore describe ontology reconstruction, not transaction-level accuracy.

Keywords: *event intelligence; FinTech; mergers and acquisitions; ontology-based semantic enrichment; SEC structured disclosures*

I. INTRODUCTION

Mergers and acquisitions remain a central mechanism through which firms respond to technological shocks, regulatory change, valuation cycles, and competitive pressure. Classic research links merger activity to industry restructuring, capital liquidity, managerial incentives, and valuation waves (Andrade et al., 2001; Harford, 2005; Malmendier & Tate, 2008; Rhodes-Kropf et al., 2005). More recent market commentary has placed artificial intelligence, private-equity exits, and capability acquisition among the forces shaping the 2025–2026 deal environment (Morgan Stanley, 2026; PricewaterhouseCoopers, 2026). These forces are particularly relevant to FinTech because financial institutions increasingly depend on payment infrastructure, lending platforms, data and cloud services, cybersecurity, regulatory technology, and AI-enabled operating capabilities.

The disclosure evidence for an M&A event is dispersed. Entry into a definitive agreement may appear under Form 8-K Item 1.01, completion of an acquisition or disposition under Item 2.01, and related financial statements or pro forma information under Item 9.01. Purchase accounting, goodwill, acquired intangible assets, contingent consideration, restructuring costs, discontinued

Received on June, 2026; Revised on July, 2026; Accepted on July, 2026; Published on August, 2026

operations, and impairment can subsequently appear in Forms 10-K and 10-Q (Securities and Exchange Commission [SEC], 2004, 2026b; SEC Investor.gov, 2021). An empirical design restricted to structured Form 8-K observations would therefore omit much of the numerical evidence and would be statistically weak in the SEC Financial Statement Data Sets, which contain only 54 structured 8-K observations during the study period. The title and research claims accordingly refer to SEC structured disclosures broadly, while treating 8-K filings as an important current-report validation layer.

Large language models and domain-specific financial language models have expanded the range of information that can be extracted from filings (Araci, 2019; Brown et al., 2020; Devlin et al., 2019; S. Wu et al., 2023; Yang et al., 2023). Retrieval-augmented and evidence-grounded systems further emphasize source attribution, numerical verification, and abstention when evidence is insufficient (Lewis et al., 2020; Li et al., 2023, 2024; Q. Wu et al., 2023; Zhong et al., 2025). The present study uses these ideas to define analyst-relevant semantic slots, but it does not run a generative model over each filing. The extraction layer is a deterministic dictionary applied to XBRL tag names, labels, and documentation. "LLM-inspired ontology-based semantic enrichment" is therefore the precise description of the method.

The study asks four questions. First, how much M&A-relevant semantic evidence is available in recent SEC structured disclosures? Second, how much predictive improvement remains when variables used to construct the labels are excluded from model inputs? Third, does performance persist across forward quarterly validation rather than one favorable split? Fourth, how well do the structured signals align with independent full-text 8-K evidence and FDIC transaction records? The contribution is not a claim that ontology-derived labels equal confirmed transactions. Instead, it is a controlled assessment of how well auditable semantic features support disclosure triage under progressively stricter validation.

The main empirical lesson is cautionary. Proximal models attain near-perfect discrimination because predictors and labels occupy closely related semantic spaces. Once all label-generating variables and their descendants are removed, performance remains useful for several tasks but decreases substantially. This gap is itself an important result because it distinguishes semantic recoverability from independent event prediction. It also motivates a practical architecture in which structured signals rank filings for analyst review, full-text evidence explains the ranking, and transaction databases provide external confirmation.

II. LITERATURE REVIEW

A. M&A event intelligence and financial disclosure

M&A research treats deal activity as the product of several related mechanisms rather than a single observable state. Andrade et al. (2001) document the scale and industrial clustering of

acquisitions; Harford (2005) links merger waves to economic shocks and liquidity; Rhodes-Kropf et al. (2005) connect merger activity to valuation dispersion; and Betton et al. (2008) synthesize evidence on takeover methods and outcomes. Malmendier and Tate (2008) show that managerial beliefs can influence acquisition choices and market reactions. These findings motivate an event-intelligence design that separates acquisition evidence, valuation signals, and post-deal integration risk.

SEC filings provide several points along that event sequence. Form 8-K supplies timely legal and current-report evidence, whereas periodic filings carry the accounting consequences of a transaction. The unit of analysis in this study is therefore a filing, not a unique completed transaction. Multiple filings may relate to one economic deal, and one periodic filing may summarize several historical acquisitions. This distinction prevents filing-level classification from being interpreted as a deal-count model and is consistent with adjacent work that separates source documents from the economic events inferred from them (Zhao & Zhang, 2024).

B. Financial NLP and evidence-grounded enrichment

Transformer models established the technical basis for contextual language processing, while BERT and GPT-style pretraining demonstrated broad transfer across downstream tasks (Brown et al., 2020; Devlin et al., 2019; Vaswani et al., 2017). FinBERT, BloombergGPT, and FinGPT show that financial vocabulary and document structure matter for domain performance (Araci, 2019; Liu et al., 2023; S. Wu et al., 2023; Yang et al., 2023). SEC-filing benchmarks likewise stress entity alignment, period tracking, and numerical reasoning (Jiang et al., 2026).

Recent financial-language-model benchmarks and disclosure studies extend this evidence to tabular reasoning, document-scale question answering, accounting narratives, and domain adaptation (Bettencourt et al., 2026; Bhatia et al., 2024; Bochkay et al., 2023; Chalkidis et al., 2022; Guo et al., 2023; Huang et al., 2023; Lee et al., 2024; Nie et al., 2024; Xie et al., 2024). Together, they support task-specific financial evaluation rather than inferring performance from a model's general language capability.

For regulated financial analysis, model output must remain connected to source evidence. Retrieval-augmented generation provides a general grounding architecture (Lewis et al., 2020), and financial applications have extended that principle to SEC/XBRL verification, risk memos, financial search, accounting disclosure review, and long-document due diligence (Chen et al., 2025; Meng et al., 2026; Q. Wu et al., 2023; Zhang et al., 2023, 2025; Zhou et al., 2023, 2024, 2026). Evidence-consistency checks and calibrated abstention further treat unsupported claims as model risk rather than a stylistic defect (Li et al., 2023, 2024; Zhong et al., 2025). These studies support a constrained enrichment layer whose categories and source terms can be audited.

Recent work provides complementary tests for such evidence controls. RAG surveys and self-reflective retrieval systems distinguish retrieval quality from generation quality, while RAGAs and ARES make that separation explicit in evaluation (Asai et al., 2024; Es et al., 2024; Gao et al., 2023; Saad-Falcon et al., 2024). Studies of parametric versus retrieved memory, hallucination detection, and atomic factual precision likewise show why a source-grounded output should be checked at the claim level (Ji et al., 2023; Mallen et al., 2023; Manakul et al., 2023; Min et al., 2023). Financial benchmarks add numerical and domain constraints to the same problem (Chen et al., 2022; Loukas et al., 2022; Lopez-Lira & Tang, 2023; Xie et al., 2023), while broad reviews clarify that the term "large language model" denotes a learned generative architecture rather than a deterministic dictionary (Zhao et al., 2023). This distinction directly motivates the terminology used in the present study.

Graph-based, corrective, and long-context retrieval studies further show that retrieval can fail through missing context, misplaced evidence, or unsupported synthesis even when the underlying source is available (Edge et al., 2024; Huang et al., 2025; Jeong et al., 2024; N. F. Liu et al., 2023; Tonmoy et al., 2024; Yan et al., 2024). The deterministic ontology avoids generated prose, but its dictionary coverage and evidence boundaries still require the same discipline.

The ontology used here resembles an analyst prompt schema—acquisition, disposition, valuation, integration, risk, and consideration cues—but replaces stochastic generation with deterministic matching. This choice provides reproducibility and provenance, but it should not be presented as an empirical test of an LLM. The method evaluates an LLM-inspired representational design.

Knowledge-graph research provides a useful conceptual bridge: controlled entities and relations can organize domain evidence even when no generative model is involved (Ji et al., 2022; Peng et al., 2023). Work connecting knowledge graphs with language models further supports explicit schemas as interfaces between unstructured language and auditable domain concepts (Pan et al., 2024). Accordingly, the ontology is treated as a versioned measurement instrument whose terms, exclusions, and descendants must be documented.

C. Explainability, leakage, and temporal evaluation

Logistic regression, decision trees, and random forests provide familiar comparisons between linear and nonlinear decision boundaries (Breiman, 2001; Pedregosa et al., 2011). Feature importance and local explanation methods are useful when model outputs must be reviewed by analysts (Lundberg & Lee, 2017; Ribeiro et al., 2016), while calibration and uncertainty analysis are central to financial model governance (Jin et al., 2024). Explainability, however, does not correct a circular target. If a label is computed from acquisition and disposition counts and those same counts are used as features, excellent performance may simply reproduce the labeling rule.

The present design addresses that problem with proximal and strict feature protocols. The proximal protocol measures how efficiently semantic enrichment reconstructs an ontology-defined target. The strict protocol measures how much predictive information remains after every label-generating input and deterministic descendant is removed. Forward quarterly evaluation further avoids random mixing of earlier and later disclosures. These safeguards make the resulting scores interpretable, although they cannot substitute for broad independent transaction labels.

This design follows evidence that leakage can materially inflate machine-learning results and that reproducible studies should report the complete split, preprocessing, and model-selection path (Kapoor & Narayanan, 2023; Kapoor et al., 2024). Rolling-origin evaluation is particularly appropriate when observations arrive chronologically and the data-generating process may shift (Hewamalage et al., 2023; Petropoulos et al., 2022; Tashman, 2000; Wiles et al., 2022). The analysis therefore reports precision, recall, and threshold-dependent error patterns rather than relying on a single aggregate metric (Hicks et al., 2022), retains dataset provenance (Pushkarna et al., 2022), and treats ontology labels as potentially noisy measurements (Song et al., 2023). Explainability is used to support review, not to validate a circular target (Bücker et al., 2022; Dwivedi et al., 2023; Longo et al., 2024; Molnar, 2022; Rudin et al., 2022).

Complementary recommendations on statistical comparison, forecasting evaluation, hyperparameter disclosure, and quantitative explainability motivate the reported fold ranges, sensitivity specifications, confidence intervals, and error taxonomy (Ali et al., 2023; Benidis et al., 2023; Bischl et al., 2023; Januschowski et al., 2022; Nauta et al., 2023; Rainio et al., 2024; Saeed & Omlin, 2023).

D. A narrower FinTech taxonomy

Sector codes alone are too broad for FinTech analysis because “financial services” includes firms without a technology-centered business model and “technology/telecom” includes firms with no financial function. The study therefore uses the product-level taxonomy in Table 2: payments, digital lending, insurtech, regtech, wealthtech and capital-markets technology, crypto infrastructure, digital-banking infrastructure, AI/data-enabled financial services, and a residual Other FinTech class. A firm must have a finance-function cue and a technology-enabled product or infrastructure cue; generic references to software, cloud, data, or telecommunications are insufficient by themselves. The taxonomy remains rule based, but it is more precise than treating entire SIC families as FinTech.

The boundary rule is consistent with research that defines FinTech through technology-enabled financial functions rather than firm-wide industry labels (Arner et al., 2016; Chen et al., 2019; Feyen et al., 2022; Gomber et al., 2017; Thakor, 2020). More specialized evidence supports

separate treatment of open-banking entrants, digital credit, lending discrimination, digital payments, and supervisory categories (Babina et al., 2025; Bartlett et al., 2022; Basel Committee on Banking Supervision, 2024; Cornelli et al., 2023; Demirgüç-Kunt et al., 2022). These distinctions make the taxonomy narrower but more defensible for sector-specific analysis.

Evidence on FinTech performance and stability also supports separating technology-enabled financial activities from conventional finance and general technology firms (Daud et al., 2022).

III. RESEARCH METHOD

A. Data sources and analytical units

The primary source is the SEC Financial Statement Data Set for each quarter from 2025Q1 through 2026Q1. Each quarterly package contains submission metadata (sub.txt), numerical facts (num.txt), tag definitions (tag.txt), and presentation relationships (pre.txt) (SEC, 2026a). The analysis joins sub.txt, num.txt, and tag.txt by accession number, retains the filing form recorded in the submission file, and aggregates numerical facts to the filing level. As Table 1 shows, the five packages contain 32,254 filings, 7,335 unique registrants, and 18,312,494 numerical facts. Although the panel contains only 54 structured Form 8-K observations, the periodic filings provide the purchase-accounting and integration evidence that follows current reports. The same table distinguishes these structured records from the 83,530 unique full-text 8-K accessions, the 1,464 Item 2.01 screening universe, the 1,800-document modeling sample, and the 192 FDIC events.

Full-text processing follows the current Form 8-K item structure, including the relationship among Items 1.01, 2.01, and 9.01 and their exhibits (SEC, 2023). XBRL taxonomy releases are versioned by filing year, so tag normalization records the applicable taxonomy and original tag text rather than assuming a fixed vocabulary (SEC, 2025a, 2025d). Filing documents and submission metadata are obtained through SEC developer resources and EDGAR interfaces (SEC, 2025b, 2025c). This separation matters because financial dictionaries and corporate-filing classifiers can be sensitive to domain-specific wording and document context (Li, 2010; Loughran & McDonald, 2011, 2016).

Table 1. Quarterly coverage of the three empirical data layers.

Quarter	Structured filings	Numerical XBRL facts	Structured 8-K	Unique EDGAR 8-K accessions	Item 2.01 screen	Text sample	FDIC events
2025Q1	6,231	3,658,551	6	16,745	238	360	24
2025Q2	7,009	3,409,930	14	18,635	256	360	39
2025Q3	6,541	3,720,081	13	16,154	321	360	39
2025Q4	6,304	3,832,977	11	16,029	322	360	36
2026Q1	6,169	3,690,955	10	15,967	327	360	54
Total	32,254	18,312,494	54	83,530	1,464	1,800	192

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 119-141
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Two supplementary sources provide evidence independent of the numerical XBRL ontology. First, quarterly SEC master indexes identify a stratified full-text sample of 1,800 Form 8-K filings. Filing documents are parsed for Items 1.01, 2.01, and 9.01 and for acquisition, merger, disposition, and exhibit language. These text indicators are not constructed from the XBRL semantic counts. Second, the FDIC merger-and-acquisition file contributes 192 bank-transaction records dated from 2025Q1 through 2026Q1. The FDIC source supplies transaction-level event evidence, although its scope is limited to insured depository institutions.

Bank transaction records are drawn from the FDIC BankFind structure-change interface and its bulk-data service, with FFIEC relationship and transformation data used as a complementary reference for name and organizational changes (Federal Deposit Insurance Corporation, 2026a, 2026b; Federal Financial Institutions Examination Council, 2026). These records remain outside model training and threshold selection.

B. FinTech taxonomy and semantic ontology

Table 2 gives the operational FinTech taxonomy. Classification is multi-label at the taxonomy stage and binary only when the modeling target is formed. Exact word boundaries, normalized names, and context terms are used to avoid substring matches. A technology cue must co-occur with a financial-function cue unless the term is intrinsically FinTech-specific, such as "payment processor" or "insurtech." This change prevents general technology, telecommunications, and conventional financial-services firms from being included solely because of a broad SIC family.

Table 2. Product-level FinTech taxonomy and observed full-text classifications.

Category	Included activities	Illustrative qualifying cues	Observed n
Payments	Merchant acquiring, card processing, wallets, transfers, settlement	payment processor, merchant acquiring, digital wallet, remittance	39
Digital lending	Digital origination, underwriting, servicing, embedded credit	digital lending, loan platform, underwriting technology, buy-now-pay-later	21
InsurTech	Digital distribution, policy administration, claims technology	insurtech, digital insurance, claims platform, policy software	3
RegTech	Compliance, identity, monitoring, reporting, fraud controls	KYC, AML technology, compliance platform, fraud analytics	0
WealthTech/capital-markets technology	Brokerage platforms, portfolio tools, market infrastructure	robo-advice, trading platform, portfolio technology	1
Crypto infrastructure	Custody, exchange, settlement, tokenization infrastructure	digital-asset custody, crypto exchange, blockchain settlement	16
Digital-banking infrastructure	Core banking, banking-as-a-service, account infrastructure	core banking platform, BaaS, digital banking infrastructure	0
AI/data-enabled financial services	AI or data products whose primary function is financial	AI underwriting, financial-risk analytics, banking data platform	0
Other FinTech	Qualifying finance-function and technology cues outside named classes	context-qualified residual cases	6
Not FinTech	No qualifying product-level combination	generic finance, software, cloud, data, or telecom cues alone	1,714

The XBRL ontology in Table 3 is applied to normalized tag names, tag labels, and tag documentation. For each category, the parser records fact counts, unique-tag counts, and absolute monetary intensity where units permit. The dictionary uses exact tokens and controlled multiword expressions; matches are case insensitive after punctuation and camel-case normalization. Negation and exclusion lists remove common false matches. Dictionary version, matched source field, tag, accession number, and category are retained for traceability.

Table 3. Ontology dictionary and filing-level feature formulas.

Category	Core dictionary concepts	Filing-level features
Acquisition (A)	business combination, acquisition, acquired, acquirer, purchase of business	fact count, unique-tag count, absolute monetary sum
Disposition (D)	disposition, divestiture, disposal group, sale of business, held for sale, discontinued operation	fact count, unique-tag count, absolute monetary sum
Valuation (V)	consideration transferred, fair value, purchase price, contingent consideration, bargain purchase	fact count, unique-tag count, absolute monetary sum
Integration (I)	goodwill, acquired intangible, customer relationship, developed technology, restructuring, synergy	fact count, unique-tag count, goodwill-plus-intangibles ratio
Risk (R)	impairment, contingent liability, litigation, regulatory, covenant, credit loss, material weakness, going concern	fact count, unique-tag count, risk intensity
Cash consideration	cash paid, cash consideration, cash acquired, net cash outflow	binary cue and count
Stock consideration	shares issued, common stock consideration, equity consideration	binary cue and count
Debt consideration	notes issued, debt assumed, borrowing for acquisition	binary cue and count

C. Filing-level preprocessing and feature construction

Duplicate accession numbers are removed before aggregation. For each accession number, the pipeline computes the number of numeric facts, unique tags, monetary facts, segmented facts, nonmissing values, total absolute numerical intensity, and maximum absolute value. Canonical financial concepts include assets, liabilities, equity, revenue, net income, operating income, cash, goodwill, intangible assets, acquisition cash paid, and business-combination consideration. Derived ratios are leverage, cash-to-assets, profit margin, and goodwill-plus-intangible-assets-to-assets. Ratios are winsorized at the training-period 1st and 99th percentiles; nonpositive-skewed intensity variables use log1p(abs(x)).

Categorical variables are one-hot encoded with unknown holdout levels ignored. Numerical variables are median-imputed using training data. Logistic regression receives standardized numerical variables; tree models receive the same imputed values without scaling. Every preprocessing operator is fitted within the training window and then applied unchanged to the next quarter, following the separation of retrieval, computation, and reporting advocated in numerical-reasoning guardrail studies (Li et al., 2026; Q. Wu et al., 2023).

D. Labels, circularity controls, and model specifications

For filing *i*, the M&A semantic intensity use eq. (1) below.

$$S_{\mathrm{i}}^{\mathrm{MA}} = A_{\mathrm{i}} + D_{\mathrm{i}} + 0.50V_{\mathrm{i}} + 0.75I_{\mathrm{i}} \qquad (1)$$

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 119-141
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

The training-period upper quartile of this score defines the principal structured M&A label; a structured 8-K also qualifies when its score exceeds the training-period 8-K median. FinTech M&A requires both the M&A label and a qualifying category from Table 2. Integration-risk-high combines risk intensity, segment complexity, and goodwill/intangible intensity. Valuation-signal-high uses the training-period median of M&A-related monetary intensity. Table 5 reports definitions and full-panel prevalence.

The proximal feature protocol excludes the outcome, the final composite score, threshold indicators, and post-label variables but retains lower-level semantic measurements. It measures semantic reconstruction. The strict protocol additionally removes every component entering the label formula, all deterministic transformations of those components, and task-specific identity fields. Thus, strict M&A models exclude acquisition, disposition, valuation, and integration counts and descendants; strict FinTech models also exclude taxonomy flags, name cues, and broad sector proxies; strict integration-risk models exclude risk counts, segment-complexity variables, and goodwill/intangible measures; and strict valuation models exclude M&A monetary-intensity variables. The 2026Q1 performance under this protocol is the principal robustness result.

Table 4 presents the complete workflow and fixed model settings. Threshold selection uses only training observations: predicted probabilities are evaluated over the training precision-recall thresholds, the threshold maximizing F1 is selected, and that value is frozen for the next-quarter test. Confidence intervals use 2,000 filing-level bootstrap resamples with random seed 42. Models and preprocessing use scikit-learn (Pedregosa et al., 2011).

The governance logic is proportionate to a decision-support model: documentation, provenance, human review, and monitoring remain necessary even when the classifier is not a generative system. This approach is consistent with recent AI risk-management standards and financial-stability assessments (Autio et al., 2024; Financial Stability Board, 2024, 2025; International Monetary Fund, 2024; International Organization for Standardization, 2023; Tabassi, 2023).

Table 4. Reproducible workflow and fixed modeling configuration.

Stage	Specification
1. Join and deduplicate	Join submission, numerical fact, and tag files by accession number; keep one filing record per accession
2. Aggregate	Compute generic counts, canonical financial concepts, ratios, and ontology-category features
3. Split	Use only earlier quarters for preprocessing, label thresholds, model fitting, and decision-threshold selection
4. Preprocess	Median numerical imputation; one-hot categorical encoding with unknown levels ignored; training-only winsorization; scaling for logistic regression
5. Logistic regression	class_weight='balanced', max_iter=2000, random_state=42; remaining scikit-learn settings at defaults
6. Decision tree	class_weight='balanced', random_state=42; remaining scikit-learn settings at defaults
7. Random forest	n_estimators=300, class_weight='balanced_subsample', n_jobs=-1, random_state=42; remaining settings at defaults
8. Decision threshold	Training probability threshold that maximizes F1, then held fixed for the forward quarter
9. Metrics	Accuracy, precision, recall, F1, ROC-AUC, average precision, and false-positive/false-negative inspection
10. Uncertainty	2,000 filing-level bootstrap resamples, seed 42, percentile 95% confidence interval

Table 5. Outcome definitions and full-panel prevalence.

Outcome	Operational definition	Positive filings	Prevalence
M&A event	Upper-quartile training M&A intensity, plus high-intensity structured 8-K rule	8,601	26.67%
FinTech M&A event	M&A event and qualifying FinTech taxonomy category	4,380	13.58%
Integration-risk-high	M&A event with training-defined high risk, segment-complexity, and goodwill/intangible intensity	6,302	19.54%
Valuation-signal-high	M&A event at or above the training median of M&A monetary intensity	4,410	13.67%

E. Temporal, text, and transaction validation

The primary split trains on 2025Q1–2025Q4 and tests on 2026Q1. To determine whether the result depends on that quarter, Table 6 defines three expanding-window tests beginning with 2025Q3. Each fold repeats preprocessing, label-threshold estimation, model fitting, and decision-threshold selection using only earlier quarters. The reported rolling means therefore summarize three genuinely forward tests rather than repeated random partitions.

Table 6. Expanding-window rolling-quarter validation design.

Training window	Training filings	Forward test quarter	Test filings	Role
2025Q1–2025Q2	13,240	2025Q3	6,541	Rolling validation
2025Q1–2025Q3	19,781	2025Q4	6,304	Rolling validation
2025Q1–2025Q4	26,085	2026Q1	6,169	Primary holdout

The full-text layer starts from 85,456 quarterly master-index rows. Removing 1,926 co-registrant duplicate rows leaves 83,530 unique accessions. An SEC full-text search identifies 1,464 Item 2.01 candidates. The 1,800-document sample contains 360 accessions per quarter, stratified as 120 screen-positive and 240 screen-negative records. Parsing identifies 120, 120, 121, 120, and 120 Item 2.01 filings across the five quarters. Models are evaluated over the same three forward quarters. A structure-only baseline uses filing metadata; a strict main-text model removes direct label terms; a strict main-text-plus-exhibit model adds exhibit-derived features while still excluding label-generating fields; and a full direct-feature model retains the direct Item 2.01 cues as a reconstruction ceiling.

FDIC event linkage normalizes organization names and searches filing dates from seven days before through ten days after the event date. Candidate links must attain name similarity of at least 0.64 and are then deduplicated by event and accession. This procedure is evaluated only as conditional positive-case sensitivity: because linkage candidates begin with Item 2.01 filings, it cannot estimate specificity or recall over every FDIC event.

Figure 1 summarizes the data boundaries. Structured filings support model development, full-text 8-K documents support event-language validation, and FDIC records support limited transaction-level validation. The three sources are connected only through explicit accession, issuer-name, and date fields; no external validation label is used to train the structured models.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 119-141
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

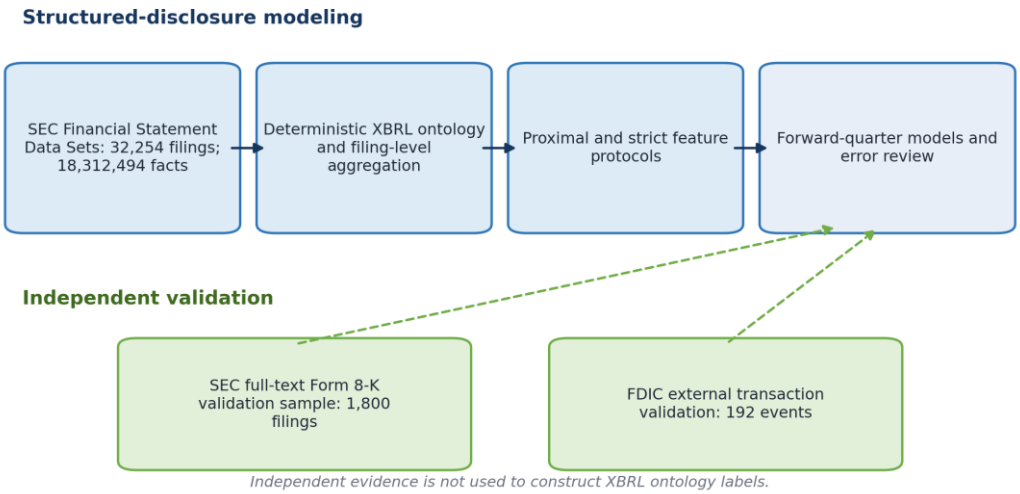

Figure 1. Data boundaries and validation workflow for structured-disclosure M&A intelligence.

IV. RESULTS AND DISCUSSION

A. Coverage and target difficulty

Table 1 shows that quarterly filing counts range from 6,169 to 7,009 and that the panel contains between 3.41 million and 3.83 million numerical facts per quarter. The form composition is dominated by 10-K and 10-Q disclosures, so the empirical claim is appropriately framed as structured-disclosure intelligence rather than an 8-K-only study. The 54 structured 8-K observations serve as current-report anchors, whereas the 1,800-record text sample provides substantially broader 8-K coverage.

Table 2 narrows the FinTech construct. General financial services and technology/telecom firms no longer qualify by sector alone. This makes FinTech M&A a more difficult target than in the broad sector definition, which helps explain its lower strict-protocol performance. Table 3 and Table 5 also show why the four tasks differ. M&A and FinTech labels are categorical triage outcomes, while integration and valuation targets summarize downstream accounting intensity. These labels identify filing characteristics; they do not establish transaction completion, deal value, or realized integration failure.

B. Circularity sensitivity in M&A detection

Table 7 reports rolling results for the structured-disclosure proxy tasks. With semantically proximal features, logistic regression attains mean M&A F1 = 0.981941 across the three forward quarters, with fold values from 0.969419 to 0.989523, ROC-AUC = 0.999517, and average precision = 0.998315. These values show that the ontology representation can reconstruct the ontology-derived target with very little error.

When every label-generating variable and deterministic descendant is excluded, mean M&A F1 decreases to 0.735640, with fold values from 0.684843 to 0.793901, ROC-AUC = 0.930870, and

average precision = 0.824107. The decline of 0.246301 F1 points is large and confirms that the proximal score cannot be interpreted as independent event-detection accuracy. FinTech M&A shows an even larger decline, from 0.921198 to 0.449503. The strict ROC-AUC remains high because it evaluates ranking, but the much lower average precision of 0.329455 reflects the difficulty of identifying the narrower positive class.

Table 7. Rolling-quarter performance for ontology-derived structured-disclosure proxies.

Task and protocol	Model	Mean F1	Fold F1 range	95% F1 CI	ROC-AUC	Average precision
M&A, proximal	Logistic regression	0.981941	0.969419–0.989523	0.979047–0.985087	0.999517	0.998315
M&A, strict	Logistic regression	0.735640	0.684843–0.793901	0.721396–0.747839	0.930870	0.824107
FinTech M&A, proximal	Random forest	0.921198	0.911392–0.926829	0.868125–0.958650	0.999705	0.974852
FinTech M&A, strict	Random forest	0.449503	0.430380–0.473684	0.327533–0.563318	0.964652	0.329455
Integration risk, strict	Random forest	0.798129	0.775777–0.826134	0.781688–0.812352	—	—
Valuation signal, strict	Random forest	0.753987	0.665526–0.832195	0.735065–0.773487	—	—

Figure 2 visualizes the proximal-to-strict F1 change for M&A and FinTech M&A. The much larger FinTech gap shows that ontology proximity and the original broad sector definition contributed strongly to the easier task. Integration-risk and valuation-signal scores are shown only under strict exclusion because those are the defensible operational estimates.

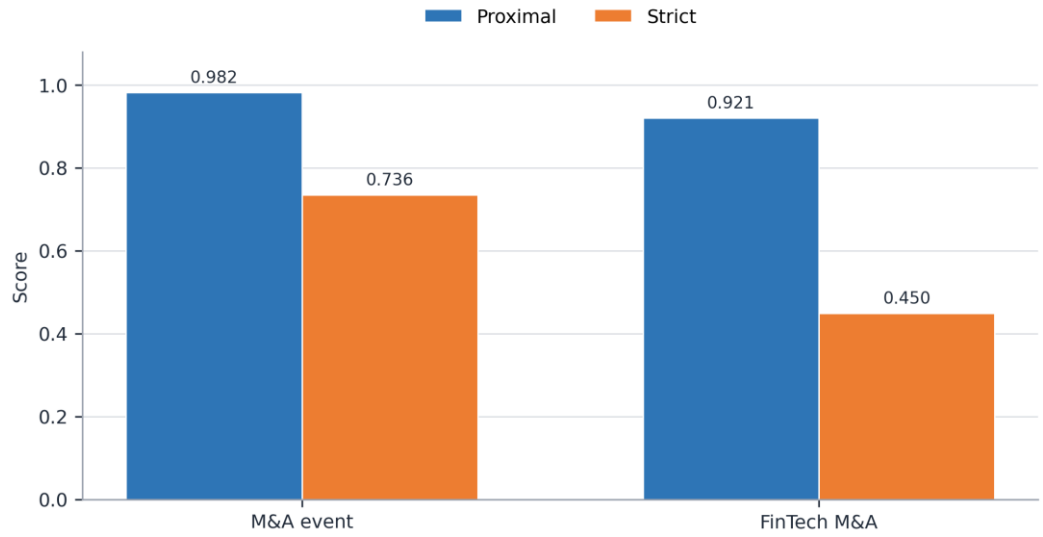

Figure 2. Proximal and strict rolling F1 for M&A and FinTech M&A proxy tasks.

C. Independent full-text Form 8-K validation

The 1,800-document text sample provides an independent evidence channel because its labels and fields are parsed from current-report items and exhibits rather than XBRL ontology counts. Table 8 shows that a structure-only decision tree reaches mean F1 = 0.217835. A strict main-text decision tree improves to 0.309927. Adding exhibit features under strict exclusion yields random-forest F1 = 0.297482, ROC-AUC = 0.873899, and average precision = 0.320178. The slight F1

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 119-141
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

decline indicates that exhibits add ranking information without automatically improving the fixed-threshold decision.

When direct Item 2.01 terms and their descendants are retained, the main-text-plus-exhibit random forest reaches F1 = 0.700555. As in the structured layer, this is a reconstruction ceiling rather than an independent estimate. The strict main-text-plus-exhibit model has precision = 0.314 and recall = 0.403, compared with precision = 0.612 and recall = 0.835 for the direct-feature model. These lower strict values are more plausible for a short, heterogeneous filing sample and reinforce the need for analyst confirmation.

Table 8. Population-weighted rolling performance for independent Item 2.01 classification.

Feature protocol	Model	Mean F1	Fold F1 range	95% F1 CI	ROC-AUC	Average precision	Prior-unseen F1
Structure baseline	Decision tree	0.217835	0.137–0.301	0.168–0.287	0.756	0.117	0.273
Strict main text	Decision tree	0.309927	0.277–0.373	0.257–0.404	0.818	0.149	0.362
Strict main text plus exhibits	Random forest	0.297482	0.268–0.354	0.246–0.367	0.874	0.320	0.332
Full direct main text plus exhibits	Random forest	0.700555	0.568–0.793	0.598–0.826	0.963	0.788	0.735

Figure 3 shows each rolling mean together with its observed fold range. The intervals make the instability of the small positive class visible and discourage interpretation of any single-quarter score as a stable deployment guarantee.

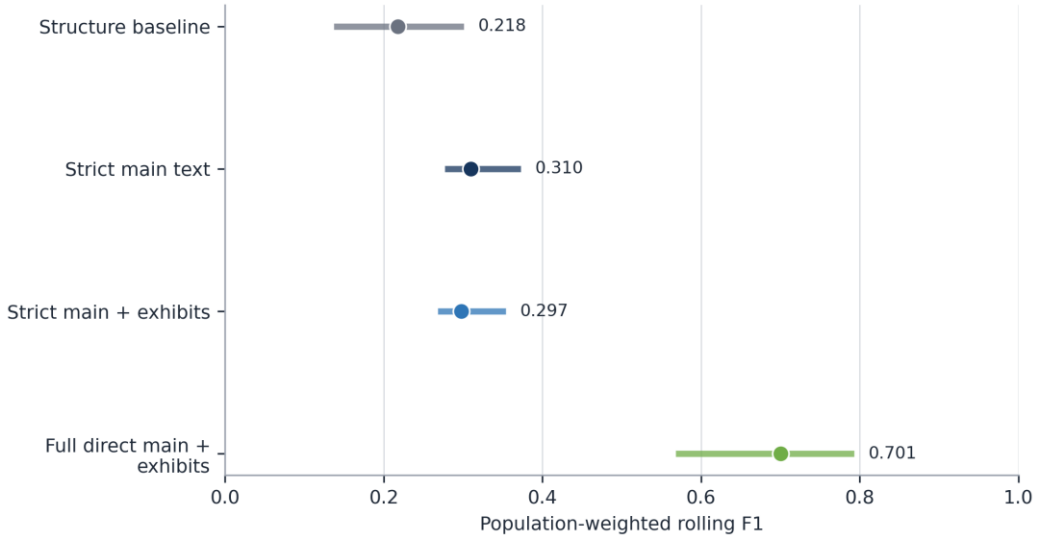

Figure 3. Rolling Item 2.01 F1 means and observed fold ranges.

D. Rolling-quarter and threshold sensitivity

The expanding windows in Table 6 require each transformation and threshold to be re-estimated without access to the forward quarter. The ordering of the models and the direction of the proximal-to-strict decline remain consistent across the three tests, while the absolute values vary as the training history expands. This pattern indicates that semantic proximity, rather than a single

2026Q1 anomaly, explains the highest scores. Three forward tests remain too few for a long-run stability claim.

Threshold sensitivity also affects operational conclusions. Lower thresholds increase recall and false positives, which may be acceptable for broad deal sourcing; higher thresholds increase precision but miss filings in which acquisition evidence is diffuse or appears mainly in exhibits. The frozen training-optimal threshold in the primary results is therefore a reproducible comparison point, not a universally optimal deployment policy.

Eighteen strict structured-label specifications vary the training quantile (70th, 75th, or 80th percentile), three score-weight systems, and whether the 8-K override is active. Their logistic-regression F1 scores range from 0.595997 to 0.759418 while mean positive prevalence ranges from 0.1730 to 0.2865. The conclusion therefore does not depend on one cutoff, although the operational class balance changes materially. In the text layer, broadening the target from strict Item 2.01 to Items 1.01 or 2.01 increases strict main-text-plus-exhibit F1 from 0.297482 to 0.396. This gain reflects an easier, broader event definition rather than better recognition of completed transactions.

E. False-positive and false-negative analysis

Strict-model errors were reviewed against the retained filing metadata, nonexcluded accounting variables, and the independent 8-K text indicators. Table 9 summarizes the observed mechanisms. False positives often arise when historical acquisition accounting remains prominent after the legal event—for example, goodwill, discontinued operations, or restructuring disclosures in a periodic report. These filings contain genuine M&A-related evidence but may not correspond to a new transaction in the test quarter.

Table 9. Error mechanisms observed in strict-protocol review.

Error direction	Filing pattern	Why the structured model can fail	Operational response
False positive	Historical purchase accounting in 10-K/10-Q	Persistent goodwill, intangibles, or contingent consideration resembles a current event	Label as post-deal accounting evidence and inspect event date
False positive	Discontinued-operation or held-for-sale accounting	Portfolio changes resemble completed acquisition/disposition evidence	Check Item 2.01 and transaction status
False positive	Restructuring or impairment without a new deal	Integration-risk signals may arise from older transactions or non-M&A stress	Separate event detection from risk monitoring
False negative	Material terms concentrated in an exhibit	Numerical XBRL facts are sparse at signing or closing	Retrieve Items 1.01/2.01/9.01 and exhibits
False negative	Amendment or abbreviated current report	The amendment may omit context present in the original filing	Link amendments to the original accession sequence
False negative	FinTech activity outside the controlled taxonomy	Narrow rules favor precision and exclude ambiguous platform firms	Route borderline taxonomy cases for analyst review

False negatives arise when decisive information is concentrated in an exhibit, when the filing has few numerical M&A concepts, when an amendment supplies only a subset of the original disclosure, or when a FinTech company falls outside the explicit taxonomy. This error structure is operationally meaningful. Some apparent false positives are useful for integration monitoring

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 2 August 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 119-141
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

even if they are not new deals, while false negatives support adding full-text exhibit retrieval. For deployment, model output should therefore be treated as a ranked review queue with an evidence reason, not an autonomous binary transaction decision.

F. Full-text and external transaction validation

Table 10 reports the FDIC linkage audit. Name/date matching produces 58 candidate events and 101 candidate filing links. Automatic deduplication retains 39 events, 55 links, and 54 unique accessions. Twenty-four linked events occur in the three rolling test quarters; 9 are detected and 15 are missed, giving conditional positive-case sensitivity of 0.375 with an exact 95% confidence interval of 0.188–0.594. Figure 4 displays the detected and missed counts.

Table 10. FDIC-to-8-K linkage and external positive-case validation.

Linkage stage	Rule or scope	Events	Filing links or accessions	Result
Candidate generation	Name similarity ≥ 0.64; filing date −7/+10 days	58	101 links	Candidate set
Automatic linkage	Deduplicated event-accession matches	39	55 links; 54 unique accessions	Linked positive cases
Rolling test subset	Linked events in 2025Q3, 2025Q4, and 2026Q1	24	—	Conditional evaluation set
Detected	Positive under the tested filing model	9	—	Conditional recall = 0.375
Missed	Not detected under the tested filing model	15	—	Exact 95% CI for recall = 0.188–0.594

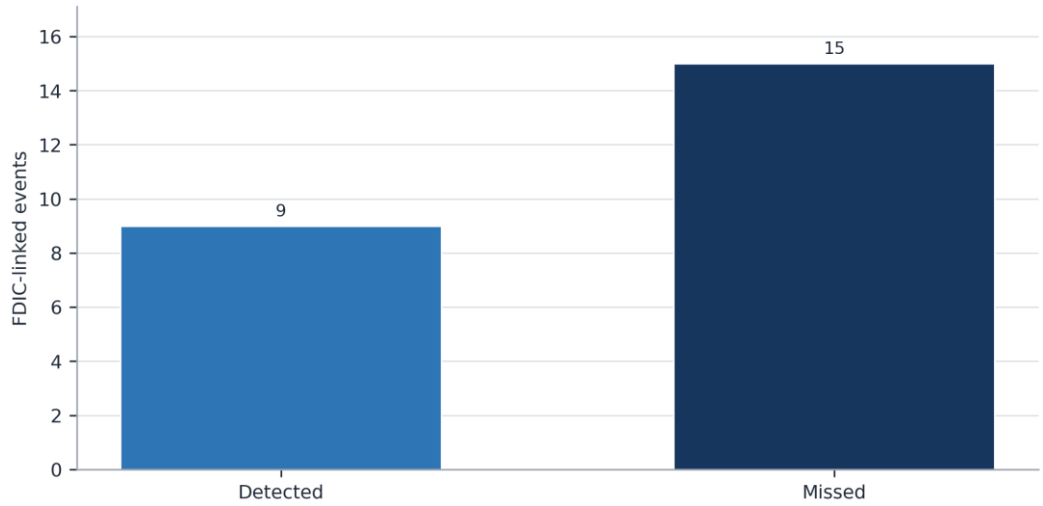

Figure 4. Detected and missed FDIC-linked events in the rolling test quarters.

The FDIC comparison is stronger than an ontology-only check because the events originate outside the SEC feature-construction process, but its conditional design must be stated precisely. All linked filings contain Item 2.01 evidence, and candidate linkage begins with that filing set. The result therefore cannot estimate specificity or coverage of all 192 FDIC events. It also cannot validate nonbank FinTech, technology vendors, private targets outside FDIC coverage, valuation intensity, or realized integration outcomes. Name normalization and bounded date windows reduce timing and entity mismatch but do not eliminate it.

G. Practical implications and limitations

The results suggest a staged workflow for corporate development, banking, private equity, and regulatory analysis. A strict model first ranks new structured disclosures; a retrieval layer then attaches exact XBRL tags and relevant 8-K passages; an analyst confirms transaction stage, parties, consideration, and risk; and an external transaction source reconciles confirmed events. Evidence cards and dashboards can expose these sources more effectively than a free-form model explanation (Li et al., 2025; Zhang et al., 2025). Rationale and consideration cues may also support financial-search reranking, provided that evidence provenance is retained (Meng et al., 2026; Su et al., 2025).

Five limitations bound the findings. First, the outcome labels remain ontology derived. Strict exclusion reduces direct circularity but does not make the labels independent of the underlying disclosure process. Second, the 1,800-filing full-text sample covers only part of the 83,530-accession universe and cannot reliably recover every scanned exhibit, image, or embedded table; material agreements may therefore remain unobserved. Third, the FDIC check is bank-only and conditional on successful Item 2.01 linkage, so it provides neither specificity nor all-event coverage. Fourth, five quarters and three forward tests provide limited evidence about recession, regulatory change, taxonomy drift, or longer M&A cycles. Fifth, the method is deterministic ontology enrichment, not direct LLM extraction, and no manually annotated market-wide gold standard is available. Future work can compare this auditable baseline with a prompt-based or fine-tuned financial language model using the same independently annotated filings.

These limitations also guide interpretation of the high proximal scores. F1 near 0.98 shows that the model can reproduce an ontology-derived filing state from semantically close variables. It does not show 98% accuracy on unique real-world transactions, deal rationales, valuation judgments, or integration outcomes. The strict and external-validation results provide the more defensible basis for claims about generalization.

V. CONCLUSION

This study presents an LLM-inspired ontology-based framework for FinTech M&A intelligence across SEC structured disclosures. The broader title reflects the actual empirical design: the principal panel is dominated by Forms 10-K and 10-Q, while Form 8-K supplies a current-report and full-text validation layer. The method converts XBRL tags into auditable acquisition, disposition, valuation, integration, risk, and consideration features and applies a narrower product-level FinTech taxonomy.

The circularity analysis changes the interpretation of the near-perfect proximal results. M&A F1 declines from 0.981941 under proximal semantic features to 0.735640 under strict exclusion, and

FinTech M&A F1 declines from 0.921198 to 0.449503. Strict integration-risk and valuation-signal models attain F1 values of 0.798129 and 0.753987. These values support filing-level triage, particularly for integration and valuation evidence, but they do not justify autonomous transaction classification.

The combination of rolling-quarter evaluation, error analysis, sampled full-text 8-K review, and 192 FDIC transactions provides a more credible validation structure than a single ontology-based holdout. The strict full-text model reaches mean F1 = 0.297482, and the FDIC audit detects 9 of 24 automatically linked rolling-test events, so external evidence is materially less favorable than the ontology-proximal scores. Broader manually annotated filings, exhaustive exhibit retrieval, nonbank transaction data, and longer temporal coverage are required before the system can be described as a general M&A event detector. The most appropriate near-term use is evidence-linked prioritization: identify filings that deserve attention, show why they were ranked, and preserve human review for transaction confirmation and economic interpretation.

REFERENCES

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion, 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Andrade, G., Mitchell, M., & Stafford, E. (2001). New evidence and perspectives on mergers. Journal of Economic Perspectives, 15(2), 103–120. https://doi.org/10.1257/jep.15.2.103

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

Arner, D. W., Barberis, J., & Buckley, R. P. (2016). The evolution of FinTech: A new post-crisis paradigm? Georgetown Journal of International Law, 47(4), 1271–1319. https://doi.org/10.2139/ssrn.2676553

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. International Conference on Learning Representations. https://openreview.net/forum?id=hSyW5go0v8

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

Babina, T., Bahaj, S., Buchak, G., De Marco, F., Foulis, A., Gornall, W., Mazzola, F., & Yu, T. (2025). Customer data access and fintech entry: Early evidence from open banking. Journal of Financial Economics, 169, 103950. https://doi.org/10.1016/j.jfineco.2024.103950

Bartlett, R. P., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. Journal of Financial Economics, 143(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047

Basel Committee on Banking Supervision. (2024). Digitalisation of finance. Bank for International Settlements. https://www.bis.org/bcbs/publ/d575.htm

Benidis, K., Rangapuram, S. S., Flunkert, V., Wang, Y., Maddix, D., Türkmen, A. C., Gasthaus, J., Bohlke-Schneider, M., Salinas, D., Stella, L., Aubet, F.-X., Callot, L., & Januschowski, T. (2023). Deep learning for time series forecasting: Tutorial and literature survey. ACM Computing Surveys, 55(6), Article 121. https://doi.org/10.1145/3533382

Bettencourt, N., Ding, X., & Giesecke, K. (2026). The Stanford EDGAR Filings Dataset: Reconstructing U.S. corporate and financial disclosures into layout-faithful and token-efficient pretraining data. arXiv. https://arxiv.org/abs/2606.18192

Betton, S., Eckbo, B. E., & Thorburn, K. S. (2008). Corporate takeovers. In B. E. Eckbo (Ed.), Handbook of empirical corporate finance (Vol. 2, pp. 291–430). Elsevier.

Bhatia, G., Nagoudi, E. M. B., Cavusoglu, H., & Abdul-Mageed, M. (2024). FinTral: A family of GPT-4 level multimodal financial large language models. arXiv. https://arxiv.org/abs/2402.10986

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. WIREs Data Mining and Knowledge Discovery, 13(2), e1484. https://doi.org/10.1002/widm.1484

Bochkay, K., Brown, S. V., Leone, A. J., & Tucker, J. W. (2023). Textual analysis in accounting: What's next? Contemporary Accounting Research, 40(2), 765–805. https://doi.org/10.1111/1911-3846.12825

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. Journal of the Operational Research Society, 73(1), 70–90. https://doi.org/10.1080/01605682.2021.1922098

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D., & Aletras, N. (2022). LexGLUE: A benchmark dataset for legal language understanding in English. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 4310–4330. https://doi.org/10.18653/v1/2022.acl-long.297

Chen, M. A., Wu, Q., & Yang, B. (2019). How valuable is FinTech innovation? The Review of Financial Studies, 32(5), 2062–2106. https://doi.org/10.1093/rfs/hhy130

Chen, Y., Zhou, S., & Lin, E. (2025, December). Accounting-aware evidence retrieval for institutional due diligence of tokenized trade receivable RWA. Journal of Technology Informatics and Engineering, 4(3), 649–663. https://doi.org/10.51903/jtie.v4i3.542

Chen, Z., Li, S., Smiley, C., Ma, Z., Shah, S., & Wang, W. Y. (2022). ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 6279–6292. https://doi.org/10.18653/v1/2022.emnlp-main.421

Cornelli, G., Frost, J., Gambacorta, L., Rau, P. R., Wardrop, R., & Ziegler, T. (2023). Fintech and big tech credit: Drivers of the growth of digital lending. Journal of Banking & Finance, 148, 106742. https://doi.org/10.1016/j.jbankfin.2022.106742

Daud, S. N. M., Ahmad, A. H., Khalid, A., & Azman-Saini, W. N. W. (2022). FinTech and financial stability: Threat or opportunity? Finance Research Letters, 47, 102667. https://doi.org/10.1016/j.frl.2021.102667

Demirgüç-Kunt, A., Klapper, L., Singer, D., & Ansar, S. (2022). The Global Findex Database 2021: Financial inclusion, digital payments, and resilience in the age of COVID-19. World Bank. https://doi.org/10.1596/978-1-4648-1897-4

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. ACM Computing Surveys, 55(9), Article 194. https://doi.org/10.1145/3561048

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., & Larson, J. (2024). From local to global: A graph RAG approach to query-focused summarization. arXiv. https://arxiv.org/abs/2404.16130

Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 150–158. https://doi.org/10.18653/v1/2024.eacl-demo.16

Federal Deposit Insurance Corporation. (2026a). BankFind Suite: Bank Structure Changes. https://banks.data.fdic.gov/bankfind-suite/oscr

Federal Deposit Insurance Corporation. (2026b). BankFind Suite: Bulk Data and API. https://banks.data.fdic.gov/bankfind-suite/bulkData

Federal Financial Institutions Examination Council. (2026). National Information Center data download: Relationships and transformations. https://www.ffiec.gov/npw/FinancialReport/DataDownload

Feyen, E., Natarajan, H., & Saal, M. (2022). Fintech and the future of finance: Market and policy implications. World Bank. https://www.worldbank.org/en/publication/fintech-and-the-future-of-finance

Financial Stability Board. (2024). The financial stability implications of artificial intelligence. https://www.fsb.org/2024/11/the-financial-stability-implications-of-artificial-intelligence/

Financial Stability Board. (2025). Monitoring adoption of artificial intelligence and related vulnerabilities in the financial sector. https://www.fsb.org/2025/10/monitoring-adoption-of-artificial-intelligence-and-related-vulnerabilities-in-the-financial-sector/

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv. https://arxiv.org/abs/2312.10997

Gomber, P., Koch, J. A., & Siering, M. (2017). Digital finance and FinTech: Current research and future research directions. Journal of Business Economics, 87, 537–580. https://doi.org/10.1007/s11573-017-0852-x

Guo, X., Xia, H., Liu, Z., Cao, H., Yang, Z., Liu, Z., Wang, S., Niu, J., Wang, C., Wang, Y., Liang, X., Huang, X., Zhu, B., Wei, Z., Chen, Y., Shen, W., & Zhang, L. (2023). FinEval: A Chinese financial domain knowledge evaluation benchmark for large language models. arXiv. https://arxiv.org/abs/2308.09975

Harford, J. (2005). What drives merger waves? Journal of Financial Economics, 77(3), 529–560. https://doi.org/10.1016/j.jfineco.2004.05.004

Hewamalage, H., Ackermann, K., & Bergmeir, C. (2023). Forecast evaluation for data scientists: Common pitfalls and best practices. Data Mining and Knowledge Discovery, 37, 788–832. https://doi.org/10.1007/s10618-022-00894-5

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. Scientific Reports, 12, 5979. https://doi.org/10.1038/s41598-022-09954-8

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. Contemporary Accounting Research, 40(2), 806–841. https://doi.org/10.1111/1911-3846.12832

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), Article 42. https://doi.org/10.1145/3703155

International Monetary Fund, Monetary and Capital Markets Department. (2024). Global Financial Stability Report, October 2024: Steadying the course—Uncertainty, artificial intelligence, and financial stability. International Monetary Fund. https://doi.org/10.5089/9798400277573.082

International Organization for Standardization. (2023). ISO/IEC 42001:2023: Information technology—Artificial intelligence—Management system. https://www.iso.org/standard/81230.html

Januschowski, T., Wang, Y., Torkkola, K., Erkkilä, T., Hasson, H., & Gasthaus, J. (2022). Forecasting with trees. International Journal of Forecasting, 38(4), 1473–1481. https://doi.org/10.1016/j.ijforecast.2021.10.004

Jeong, S., Baek, J., Cho, S., Hwang, S. J., & Park, J. (2024). Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 7036–7050. https://doi.org/10.18653/v1/2024.naacl-long.389

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2022). A survey on knowledge graphs: Representation, acquisition, and applications. IEEE Transactions on Neural Networks and Learning Systems, 33(2), 494–514. https://doi.org/10.1109/TNNLS.2021.3070843

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), Article 248. https://doi.org/10.1145/3571730

Jiang, Y., Chen, J., Makri, E., Chen, J., Li, P., Maatouk, A., Tassiulas, L., Brenner, E., Xiang, B., & Ying, R. (2026). Fin-RATE: A real-world financial analytics and tracking evaluation benchmark for LLMs on SEC filings. arXiv. https://arxiv.org/abs/2602.07294

Jin, J., Huang, T., & Lu, S. (2024, June). A model-risk-friendly probability of default workflow: Calibration, distribution-free uncertainty quantification, and SHAP explanations on the UCI credit card default dataset. Journal of Advanced Computing Systems, 4(6), 74–85. https://doi.org/10.69987/JACS.2024.40606

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Kapoor, S., Cantrell, E. M., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E., Hofman, J. M., Hullman, J., Lones, M. A., Malik, M. M., Nanayakkara, P., Poldrack, R. A., Raji, I. D., Roberts, M., Salganik, M. J., Serra-Garcia, M., Stewart, B. M., Vandewiele, G., & Narayanan, A. (2024). REFORMS: Consensus-based recommendations for machine-learning-based science. Science Advances, 10(18), eadk3452. https://doi.org/10.1126/sciadv.adk3452

Lee, J., Stevens, N., Han, S. C., & Song, M. (2024). A survey of large language models in finance (FinLLMs). arXiv. https://arxiv.org/abs/2402.02315

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

Li, C., Bai, J., & Wang, S. (2024, February). Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications. Journal of Advanced Computing Systems, 4(2), 76–92. https://doi.org/10.69987/JACS.2024.40207

Li, C., Su, W., & Zhang, E. (2023, January). Lightweight hallucination firewall for enterprise LLM applications: Evidence consistency, self-checking, and small-model detection on TruthfulQA. Journal of Advanced Computing Systems, 3(1), 49–65. https://doi.org/10.69987/JACS.2023.30104

Li, F. (2010). The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. Journal of Accounting Research, 48(5), 1049–1102. https://doi.org/10.1111/j.1475-679X.2010.00382.x

Li, Z., Zhang, K., & Wong, A. (2026, June). Numerical-reasoning guardrails for a quant research assistant: A compact reproducible benchmark using SEC and FRED data. Journal of Technology Informatics and Engineering, 5(2), 75–90. https://doi.org/10.51903/jtie.v5i2.541

Li, Z., Zhou, S., & Zhou, Z. (2025, May). Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench. International Journal of Graphic Design, 3(1), 196–210. https://doi.org/10.51903/ijgd.v3i1.3715

Liu, N. F., Zhang, T., & Liang, P. (2023). Evaluating verifiability in generative search engines. Findings of the Association for Computational Linguistics: EMNLP 2023, 7001–7025. https://doi.org/10.18653/v1/2023.findings-emnlp.467

Liu, X.-Y., Wang, G., Yang, H., & Zha, D. (2023). FinGPT: Democratizing Internet-scale data for financial large language models. arXiv. https://arxiv.org/abs/2307.10485

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Paez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion, 106, 102301. https://doi.org/10.1016/j.inffus.2024.102301

Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. arXiv. https://arxiv.org/abs/2304.07619

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. Journal of Accounting Research, 54(4), 1187–1230. https://doi.org/10.1111/1475-679X.12123

Loukas, L., Fergadiotis, M., Chalkidis, I., Spyropoulou, E., Malakasiotis, P., Androutsopoulos, I., & Paliouras, G. (2022). FiNER: Financial numeric entity recognition for XBRL tagging. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 4419–4431. https://doi.org/10.18653/v1/2022.acl-long.303

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 9802–9822. https://doi.org/10.18653/v1/2023.acl-long.546

Malmendier, U., & Tate, G. (2008). Who makes acquisitions? CEO overconfidence and the market's reaction. Journal of Financial Economics, 89(1), 20–43. https://doi.org/10.1016/j.jfineco.2007.07.002

Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 9004–9017. https://doi.org/10.18653/v1/2023.emnlp-main.557

Meng, S., Chen, J., & Zheng, I. (2026, April). LLM-inspired offline reranking for financial search: Query rewriting, hybrid retrieval, and listwise relevance ranking on FiQA. Journal of Technology Informatics and Engineering, 5(1), 361–378. https://doi.org/10.51903/jtie.v5i1.537

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long-form text generation. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 12076–12100. https://doi.org/10.18653/v1/2023.emnlp-main.741

Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Morgan Stanley. (2026). 5 forces driving M&A in 2026. https://www.morganstanley.com/insights/articles/mergers-and-acquisitions-outlook-2026-activity

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. ACM Computing Surveys, 55(13s), Article 295. https://doi.org/10.1145/3583558

Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024). A survey of large language models for financial applications: Progress, prospects and challenges. arXiv. https://arxiv.org/abs/2406.11903

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering, 36(7), 3580–3599. https://doi.org/10.1109/TKDE.2024.3352100

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Peng, C., Xia, F., Naseriparsa, M., & Osborne, F. (2023). Knowledge graphs: Opportunities and challenges. Artificial Intelligence Review, 56(11), 13071–13102. https://doi.org/10.1007/s10462-023-10465-9

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., ... Ziel, F. (2022). Forecasting: Theory and practice. International Journal of Forecasting, 38(3), 705–871. https://doi.org/10.1016/j.ijforecast.2021.11.001

PricewaterhouseCoopers. (2026). Global M&A industry trends: 2026 outlook. https://www.pwc.com/gx/en/services/deals/trends.html

Pushkarna, M., Zaldivar, A., & Kjartansson, O. (2022). Data cards: Purposeful and transparent dataset documentation for responsible AI. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1776–1826. https://doi.org/10.1145/3531146.3533231

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. Scientific Reports, 14, 6086. https://doi.org/10.1038/s41598-024-56706-x

Rhodes-Kropf, M., Robinson, D. T., & Viswanathan, S. (2005). Valuation waves and merger activity: The empirical evidence. Journal of Financial Economics, 77(3), 561–603. https://doi.org/10.1016/j.jfineco.2004.06.015

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys, 16, 1–85. https://doi.org/10.1214/21-SS133

Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 338–354. https://doi.org/10.18653/v1/2024.naacl-long.20

Saeed, W., & Omlin, C. W. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems, 263, 110273. https://doi.org/10.1016/j.knosys.2023.110273

Securities and Exchange Commission Investor.gov. (2021). How to read an 8-K. https://www.investor.gov/introduction-investing/general-resources/news-alerts/alerts-bulletins/how-read-8

Securities and Exchange Commission. (2004). Additional Form 8-K disclosure requirements and acceleration of filing date. https://www.sec.gov/rules-regulations/2004/03/additional-form-8-k-disclosure-requirements-acceleration-filing-date

Securities and Exchange Commission. (2023). Form 8-K: Current report pursuant to Section 13 or 15(d) of the Securities Exchange Act of 1934. https://www.sec.gov/files/form8-k.pdf

Securities and Exchange Commission. (2025a). 2025 XBRL taxonomies update. https://www.sec.gov/newsroom/whats-new/2503-2025-xbrl-taxonomies-update

Securities and Exchange Commission. (2025b). Developer resources. https://www.sec.gov/about/developer-resources

Securities and Exchange Commission. (2025c). EDGAR application programming interfaces (APIs). https://www.sec.gov/search-filings/edgar-application-programming-interfaces

Securities and Exchange Commission. (2025d). EDGAR Release 25.1. https://www.sec.gov/submit-filings/edgar-news-announcements/edgar-release-25-1

Securities and Exchange Commission. (2026a). Financial Statement Data Sets. https://www.sec.gov/data-research/sec-markets-data/financial-statement-data-sets

Securities and Exchange Commission. (2026b). Exchange Act Form 8-K: Compliance and disclosure interpretations. https://www.sec.gov/rules-regulations/staff-guidance/compliance-disclosure-interpretations/exchange-act-form-8-k

Song, H., Kim, M., Park, D., Shin, Y., & Lee, J.-G. (2023). Learning from noisy labels with deep neural networks: A survey. IEEE Transactions on Neural Networks and Learning Systems, 34(11), 8135–8153. https://doi.org/10.1109/TNNLS.2022.3152527

Su, W., Chen, S., & Zhao, C. (2025, December). Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG benchmark. Journal of Technology Informatics and Engineering, 4(3), 649–662. https://doi.org/10.51903/jtie.v4i3.543

Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. International Journal of Forecasting, 16(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0

Thakor, A. V. (2020). Fintech and banking: What do we know? Journal of Financial Intermediation, 41, 100833. https://doi.org/10.1016/j.jfi.2019.100833

Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. arXiv. https://arxiv.org/abs/2401.01313

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

Wiles, O., Gowal, S., Stimberg, F., Rebuffi, S.-A., Ktena, I., Dvijotham, K., & Cemgil, A. T. (2022). A fine-grained analysis on distribution shift. International Conference on Learning Representations. https://openreview.net/forum?id=Dl4LetuLdyK

Wu, Q., Bai, J., & Zhou, X. (2023, March). Evidence-grounded financial RAG: Reducing numerical hallucination in LLM-generated corporate risk memos. Journal of Advanced Computing Systems, 3(3), 65–84. https://doi.org/10.69987/JACS.2023.30306

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. https://arxiv.org/abs/2303.17564

Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., Xiao, M., Li, D., Dai, Y., Feng, D., Xu, Y., Kang, H., Kuang, Z., Yuan, C., Yang, K., Luo, Z., Zhang, T., Liu, Z., Xiong, G., ... Huang, J. (2024). FinBen: A holistic financial benchmark for large language models. arXiv. https://arxiv.org/abs/2402.12659

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

Yan, S.-Q., Gu, J.-C., Zhu, Y., & Ling, Z.-H. (2024). Corrective retrieval augmented generation. arXiv. https://arxiv.org/abs/2401.15884

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

Zhang, K., Chen, Y., & Qian, A. (2025, October). Evidence-grounded accounting disclosure review cards: A visual communication framework for LLM-style explanations over SEC financial statements and notes. International Journal of Graphic Design, 3(2), 395. https://doi.org/10.51903/ijgd.v3i2.3710

Zhang, K., Meng, S., & Zhou, E. (2023, February). Evidence-grounded trading desk risk memos over SEC filings: Retrieval-augmented generation with XBRL numeric verification. Journal of Advanced Computing Systems, 3(2), 60–76. https://doi.org/10.69987/JACS.2023.30205

Zhao, G., & Zhang, K. (2024, March). From earnings calls to earnings events: An LLM-guided separation of discretionary and systematic volatility trading signals. Journal of Advanced Computing Systems, 4(3), 126–143. https://doi.org/10.69987/JACS.2024.40309

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2023). A survey of large language models. arXiv. https://arxiv.org/abs/2303.18223

Zhong, Z. S., Chen, J., Zhong, E., & Sun, X. (2025, August). Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0. Journal of Technology Informatics and Engineering, 4(2), 502–520. https://doi.org/10.51903/jtie.v4i2.536

Zhou, S., Chen, Y., & Lee, K. (2026, June). Accounting-aware evidence-constrained agents for disclosure, settlement, and secondary-market risk monitoring in tokenized RWA infrastructure. Journal of Technology Informatics and Engineering, 5(2), 60–74. https://doi.org/10.51903/jtie.v5i2.544

Zhou, S., Li, Z., & Wang, E. (2023, July). Evidence-grounded RAG for tokenized trade receivable disclosure QA under U.S. capital market standards. Journal of Advanced Computing Systems, 3(7), 41–57. https://doi.org/10.69987/JACS.2023.30704

Zhou, S., Li, Z., & Wang, E. (2024, August). Long-document RAG for contractual and insurance clause analysis in receivables RWA structures. Journal of Advanced Computing Systems, 4(8), 88–104. https://doi.org/10.69987/JACS.2024.40810