FinReporting: An Agentic Workflow for Localized Reporting of Cross-Jurisdiction Financial Disclosures

Fan Zhang[1,2*] **Mingzi Song**[3] **Rania Elbadry**[1] **Yankai Chen**[1,4†] **Shaobo Wang**[2] **Yixi Zhou**[1] **Xunwen Zheng**[5] **Yueru He**[6] **Yuyang Dai**[7] **Georgi Georgiev**[8] **Ayesha Gull**[9] **Muhammad Usman Safder**[9] **Fan Wu**[1] **Liyuan Meng**[1] **Fengxian Ji**[1] **Junning Zhao**[2] **Xueqing Peng**[10*] **Jimin Huang**[10] **Yu Chen**[2] **Xue (Steve) Liu**[1,4] **Preslav Nakov**[1] **Zhuohan Xie**[1]

[1]MBZUAI [2]The University of Tokyo [3]Meiji Gakuin University [4]McGill University [5]Kyoto University [6]Columbia University [7]University of California, Berkeley [8]Sofia University "St. Kliment Ohridski" [9]Namal University [10]The Fin AI

Abstract

Financial reporting systems increasingly leverage Large Language Models (LLMs) to extract and summarize corporate disclosures. However, most existing approaches assume a single-market setting and overlook structural differences across jurisdictions. Variations in accounting taxonomies, tagging infrastructures (e.g., XBRL vs. PDF), and aggregation conventions introduce substantial challenges for semantic alignment and reliable verification. Here, we aim to bridge this gap. We present FinReporting, an agentic workflow for localized cross-jurisdiction financial reporting. The system constructs a unified canonical ontology spanning the income statement, balance sheet, and cash flow statement, and decomposes reporting into auditable stages, including filing acquisition, extraction, canonical mapping, and anomaly logging. Rather than treating LLMs as free-form generators, FinReporting employs them as constrained verifiers operating under explicit decision rules with evidence grounding. Evaluated on annual filings from the USA, Japan, and China, FinReporting improves consistency and reliability under heterogeneous reporting regimes. We further release an interactive demo that enables cross-market inspection and supports structured export of localized financial statements. Our demo is available at https://huggingface.co/spaces/BoomQ/FinReporting-Demo. A video describing our system is available at https://www.youtube.com/watch?v=f65jdEL31Kk.

1 Introduction

Financial statements are indispensable for investors to assess a firm's financial condition and performance. With recent breakthroughs in Large Language Models (LLMs), a growing line of work has begun to automate customized financial reporting.

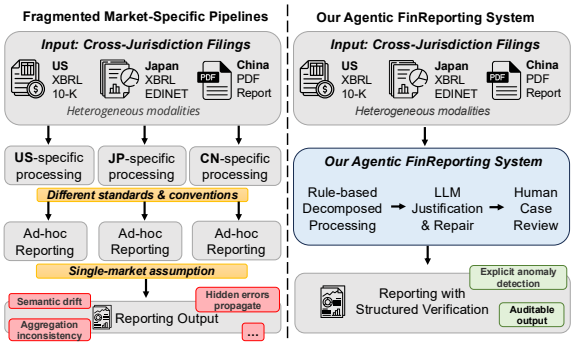

Figure 1: Financial reporting is traditionally handled via separate pipelines operating under the *single-market* assumption, leading to implicit issues. Our FinReporting system specifically implements the localized reporting of cross-jurisdiction financial disclosures.

Given a user's analytical focus, systems parse long filings together with relevant tables, and extract or summarize the financial facts that matter most to that user (Huang et al., 2024; Aguda et al., 2024; Wang and Brorsson, 2025; Shu et al., 2025; Wang et al., 2025). This reduces the burden of reading complete reports, and makes financial disclosures more accessible for decision-making.

However, existing systems predominantly operate under a *single-market assumption*: the user queries filings within the same jurisdiction where they are already familiar with the accounting standards, disclosure conventions, and reporting requirements, as shown in the left-hand side of Figure 1. For global investors, this assumption breaks. When an investor attempts to understand a foreign firm's financial statements, two practical frictions arise. On the one hand, national financial infrastructures, e.g., auditing regimes, taxonomy standards, and consolidation conventions, can vary substantially across jurisdictions. Thus, investors who are not deeply familiar with the target market often struggle to interpret raw filings correctly.

*zhang-fan@g.ecc.u-tokyo.ac.jp
†corresponding author: yankaichen@acm.org

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 728–735

On the other hand, financial data vendors, such as Bloomberg[1] and Wind,[2] rely heavily on manual extraction and expert validation to curate structured financial data. Although this ensures high accuracy, it is labor-intensive, costly, and difficult to scale, leading to high subscription costs and limited accessibility, particularly for small and mid-sized investors. Therefore, these realities motivate the need for an automatic framework that can *restructure cross-jurisdictional financial data into a localized representation aligned with the investor's home-market logic*.

However, cross-jurisdictional reporting localization extends beyond translation or format conversion. Divergent accounting taxonomies and aggregation conventions mean that superficially similar line items may represent different concepts, while equivalent concepts may appear under different labels, leading to semantic drift and inconsistent roll-ups. Because such misalignments can remain hidden and propagate into downstream quantitative models, reliable localization must explicitly support structured verification and repair rather than treating filings as visual document parsing.

To fill this gap, we propose FinReporting, an agentic workflow for unified, localized reporting of cross-jurisdiction financial disclosures. FinReporting constructs a canonical financial ontology based on universal reporting standards, e.g., Income Statement (IS), Balance Sheet (BS), and Cash Flow (CF), to align semantically equivalent items across markets. In this work, we implement this ontology across United States (US), Japanese (JP), and Chinese (CN) markets. Operationally, FinReporting decomposes end-to-end localization into a sequence of auditable steps before *Outputting*, including *Filing Acquisition*, *Statement Identification*, *Extraction*, *Canonical Mapping*. Moreover, FinReporting includes a structured verification mechanism between steps. Specifically, at each step, it combines rule-based constraints with LLM-based reasoning to validate, repair, and justify intermediate outputs, ensuring that resulting localized statements are both logically coherent and semantically faithful to source filings. Therefore, this enables FinReporting to go beyond separate information processing and surface anomalies in information aggregation for financial reporting, as shown in the right-hand side of Figure 1.

[1] https://www.bloomberg.com
[2] https://www.wind.com.cn

As illustrated by the demo interface in Figure 2, FinReporting enables users to load market-specific filings by symbol, view extracted IS/BS/CF statements, and export localized financial reports. It provides an auditable infrastructure for downstream applications, including financial question answering, regulatory analysis, and cross-market benchmarking. By operationalizing structured verification and canonical alignment, it enables transparent and consistent cross-jurisdiction financial reporting. Our contributions can be summarized as follows:

- We present FinReporting, a system for localized reporting of cross-jurisdiction financial disclosures across heterogeneous reporting standards and formats, enabling unified canonical alignment and consistent cross-market interpretation.
- We implement an auditable agentic workflow that integrates rule-based extraction, ontology-guided canonicalization, and constrained LLM-based verification and repair, producing structured outputs with explicit audit trails, anomaly flags, and quality signals (see Figure 3).
- We develop an interactive interface that allows users to explore localized financial statements, inspect verification evidence and audit trails, and export structured reports for transparent analysis and downstream applications.

2 Related Work

2.1 Financial NLP

Financial NLP has long studied textual signals in corporate disclosures, such as sentiment, tone, and linguistic cues for market prediction and risk assessment (Loughran and McDonald, 2011; Kogan et al., 2009; Qian et al., 2025a; Zhou et al., 2026). More recently, the field has expanded toward numerical reasoning and question answering over financial reports, with benchmarks such as FinQA, TAT-QA, and conversational extensions (Chen et al., 2021; Zhu et al., 2021; Chen et al., 2022; Xie et al., 2026; Qian et al., 2025b; Ji et al., 2025), as well as retrieval-centric long-form settings, e.g., FinTextQA (Chen et al., 2024). In parallel, finance-oriented foundation models and evaluation suites, e.g., BloombergGPT, FinGPT, FinBen, MultiFinBen, have been proposed to strengthen domain adaptation and capability measurement (Wu et al., 2023; Yang et al., 2023; Xie et al., 2024; Peng et al., 2025).

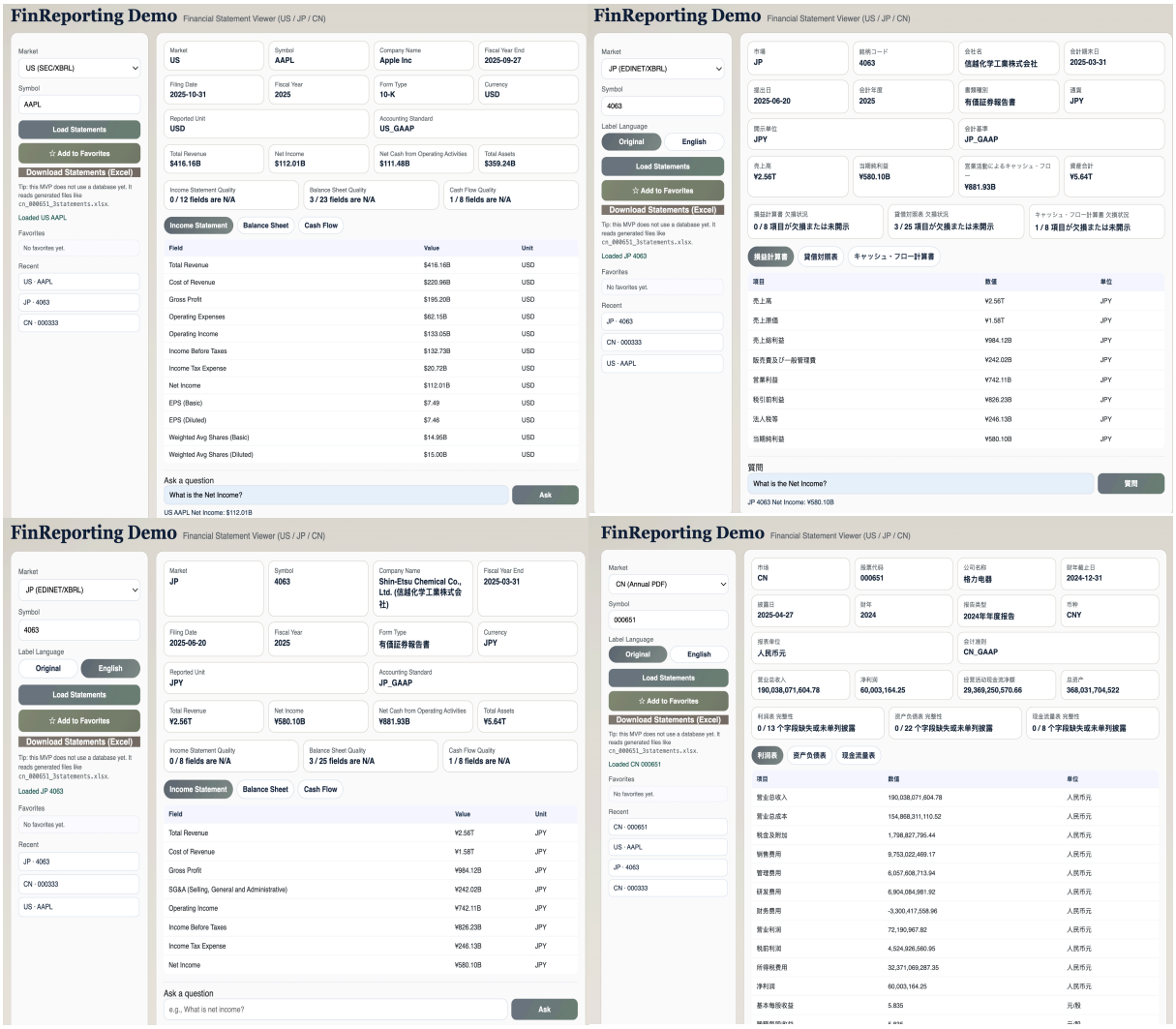

Figure 2: The FinReporting demo interface: the system supports cross-jurisdiction statement browsing and reporting (US/JP/CN), with specific interested fields selected.

2.2 Financial Disclosure Reporting

Modern financial disclosure systems increasingly adopt XBRL, i.e., standardized digital tags for financial data, to enable machine-readable reporting. For instance, the U.S. SEC provides large-scale structured datasets derived from these tags, allowing direct access to reported facts (U.S. Securities and Exchange Commission, 2025a,b). Building on such infrastructures, recent research has proposed pipelines for extracting and structuring tagged financial facts, including LLM-oriented end-to-end benchmarks for financial information extraction and structuring (Wang et al., 2025) and instruction-tuned models for taxonomy-scale extreme classification (Khatuya, 2024). Recent LLM-based interfaces for querying XBRL-tagged disclosures, e.g., XBRL-centered analysis agents, further improve usability (Han et al., 2024).

However, disclosure practices and reporting standards vary substantially across markets and regulatory environments worldwide. Although many jurisdictions have introduced digital reporting mechanisms, their taxonomies and reporting conventions are not directly interchangeable (IFRS Foundation, 2024; European Securities and Markets Authority, 2025; Ji et al., 2026). As a result, most existing pipelines are based on the *single-market* setting, as they assume a fixed taxonomy and reporting logic, limiting scalability and generalization across jurisdictions. Unlike existing XBRL-centric or single-market extraction pipelines that assume fixed taxonomies and homogeneous infrastructures, FinReporting explicitly models cross-jurisdiction heterogeneity and embeds structured verification, thus enabling robust, scalable, and auditable localization across heterogeneous filing regimes.

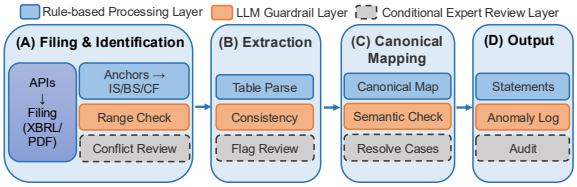

Figure 3: Overview of the FinReporting processes.

3 FinReporting System

FinReporting is an agentic workflow that partitions financial reporting into auditable steps: *Filing Acquisition*, *Statement Identification*, *Extraction*, and *Canonical Mapping*, followed by the final *Outputting* step. Between steps, FinReporting inserts structured verification mechanisms, i.e., LLM justification and human-expert review, to ensure intermediate output quality. Thus, as shown in Figure 3, the system adopts a three-layer design to sequentially execute these steps: ❶ a deterministic rule-based processing layer that produces reproducible initial results, ❷ an LLM verification/repair layer with strict guardrails, and ❸ a lightweight human review and evaluation layer for high-impact cases.

3.1 Rule-based Processing Layer

Filing Acquisition and Statement Identification. FinReporting first acquires the target annual filing and identifies the relevant statement sections with the given query. The acquisition strategy is jurisdiction-aware: for XBRL-native (US/JP) markets, FinReporting directly loads tagged facts from the filing; for PDF-centric (CN) markets, it locates the corresponding annual report and detects the page ranges of the core statements such as IS/BS/CF. This step identifies a statement-level context package (including document pointers, period metadata, and statement boundaries) that serves as the shared input to downstream extraction.

Extraction. FinReporting then proceeds with rule-based parsers to generate stable, explainable, and reproducible candidate values without relying on LLM generation. We distinguish two extraction tracks because the input regimes are fundamentally different. In XBRL-native jurisdictions (US/JP), filings already provide standardized, machine-readable tags; extraction mainly reduces to *selecting the correct facts* under reporting context, e.g., consolidated vs separate, period length, and instant/duration matching, and exporting the tagged items into our statement schema.

In PDF-centric jurisdictions (CN), disclosures are not consistently tagged and table layouts vary across issuers. Therefore, extraction requires *document-level decomposition*, table parsing, robust column selection with fallbacks, and per-field status labeling. In this work, we additionally attach a per-field status label to each extracted item to explicitly surface uncertainty, such as `OK`, `MISSING`, `PARSE_ERROR`, and `NOT_APPLICABLE`.

Canonical Mapping via a Global Ontology. In order to localize statements into an investor-home representation, FinReporting maps the extracted items into a unified canonical schema guided by a global financial ontology spanning IS/BS/CF. The ontology defines a set of core concepts and their cross-market correspondences, enabling semantically aligned structuring across heterogeneous filings. Canonical mapping produces localized statement tables under the same concept inventory, which makes cross-jurisdiction comparison and downstream applications (e.g., QA and benchmarking) consistent.

Outputting. Finally, FinReporting outputs (*i*) localized financial statements under the unified canonical schema, and (*ii*) an anomaly log and audit trail that record the irregularities and decisions encountered in previous steps.

3.2 LLM Guardrail Layer

To prevent *undetected mistakes*, i.e., plausible-looking outputs that are incorrect but not flagged, FinReporting treats the LLM as a *bounded verifier* rather than a free-form extractor. At each stage, FinReporting combines pre-defined constraints with LLM-based reasoning to (*i*) validate intermediate outputs, (*ii*) propose repairs when evidence is sufficient, and (*iii*) justify decisions with traceable support. The verifier operates under a restricted decision space: `KEEP` (retain the rule value), `REPAIR` (override with an evidence-backed value), or `NEED_REVIEW` (defer to human). A repair is applied only when all guardrails pass: the field is repairable (typically `MISSING` or `PARSE_ERROR`), the evidence is explicitly grounded in the filing context, and the proposed value is consistent with the cited evidence; otherwise, the system falls back to `NEED_REVIEW`. All decisions are logged with evidence and failure reasons, ensuring that the final localized statements remain logically coherent and semantically faithful to the source filings.

Jurisdiction	Metric	$\text{LLM}_{\text{Reporting}}$	FinReporting
US-Jurisdiction	FR	94.44	95.56
	CR	5.56	15.56
	ACC	89.38	90.23
JP-Jurisdiction	FR	84.44	84.44
	CR	15.56	15.56
	ACC	88.36	88.36
CN-Jurisdiction	FR	63.33	63.33
	CR	26.67	40.56
	ACC	78.15	82.11

Table 1: Performance across jurisdictions. In this experiment, we use `GPT-4o` in our LLM guardrail layer.

3.3 Conditional Expert Review Layer

FinReporting supports targeted human review for cases flagged as `NEED_REVIEW` or large discrepancies. Reviewers inspect audit trails and evidence to resolve conflicts. For demonstration, we provide structured templates and ablation comparisons to quantify how verification and repair affect extraction completeness and localization consistency.

4 Experiments

4.1 Setup

Evaluation Protocol. We evaluate FinReporting over core financial statement fields from annual filings. For each company, the system produces reporting for a fixed inventory of eighteen core items spanning IS/BS/CF, with an auditable trace and a review flag when evidence is insufficient.

Evaluation Data. We evaluate FinReporting under a unified annual-filing protocol across the US, JP, and CN, focusing on non-financial firms (excluding banks, insurers, and securities firms due to different reporting schemas), consolidated statements, and canonical IS/BS/CF mapping. Data are collected from public regulatory disclosures: US filings from SEC EDGAR (10-K XBRL),[3] JP filings from EDINET annual securities reports (XBRL),[4] and CN annual reports from publicly disclosed PDFs.[5] For each market, we construct a 20-company baseline split for rule development and gold annotation, and a 10-company challenge split as a held-out evaluation set, yielding 30 firms per market (90 total). Gold annotations are manually validated by financial experts to ensure correct canonical mappings and numerical consistency.

[3] `https://www.sec.gov/search-filings`
[4] `https://disclosure2.edinet-fsa.go.jp/`
[5] `http://www.cninfo.com.cn/`

LLM Backbones	FR	CR	ACC	Cost ($)
`GPT-5.2`	95.56	8.89	90.23	36.96
`GPT-5 mini`	95.56	15.00	90.23	17.77
`GPT-4o`	95.56	15.56	90.00	34.04
`Gemini-2.5-Flash`	95.56	12.78	90.23	7.27
`Gemini-2.5-Flash-Lite`	95.56	8.89	90.00	1.47
`DeepSeek-Chat`	95.56	100.00	90.23	2.41

Table 2: LLM backbone comparison on US filings.

Cross-market metrics are reported on a shared subset of 18 canonical targets (IS: 5, BS: 7, CF: 6), while market-specific items are retained for UI display and qualitative analysis.

Evaluation Metrics. Let N be the total number of target fields. We report (*i*) *Filled Rate (FR)*, the fraction of fields with non-null outputs, (*ii*) *Conflict Rate (CR)*, the fraction of fields triggering human review due to disagreement between the deterministic result and the LLM verifier (or insufficient evidence), and (*iii*) *Accuracy (Acc)*, computed against manual labels over the reviewed fields.

4.2 Empirical Analysis

Comparison to Naïve LLM Reporting Pipeline. As shown in Table 1, the performance is strongest for US examples, slightly lower for Japanese, and substantially weaker for Chinese for both methods. This pattern is largely driven by differences in data structure and standardization. US disclosures are highly machine-readable and schema-consistent (e.g., standardized tags and stable semantics), making the task closer to structured extraction and mapping; Japan follows a similar paradigm, but shows greater variability in tagging and reporting, increasing cross-firm alignment complexity. In contrast, China often requires extracting and reconstructing information from PDF reports, where layout variability, table fragmentation, and inconsistent line items introduce noise and amplify error propagation, resulting in the lowest reliability.

Deployment of Backbone LLMs. Table 2 compares different LLMs in the US Jurisdiction setting. We can see that the overall fill rate is consistent across models, which suggests that coverage is driven primarily by the task pipeline rather than the choice of a backbone. From a practical deployment perspective, the results indicate that smaller/efficient backbones can match the strongest models in terms of accuracy while being much cheaper, making them attractive default choices for real-world environments.

5 The Demo Application

Overview We develop a lightweight web-based demo for interactive exploration of financial statements extracted by FinReporting. It targets financial analysts, cross-market investors, and financial NLP researchers requiring structured, auditable cross-jurisdiction reporting. Users select a market (US, JP, CN) and company, and the interface renders the three canonical statements—Income Statement (IS), Balance Sheet (BS), and Cash Flow (CF)—in tabbed views. Each view displays metadata such as fiscal year, filing date, currency, and accounting standard, and highlights key indicators including revenue, net income, total assets, and net operating cash flow.

Unified Schema and QA The outputs are normalized under a unified schema across markets while preserving market-specific labels and units. This enables consistent comparison without discarding local reporting semantics. To support common analyst queries, the demo provides template-based question answering for high-frequency metrics, including revenue, net income, and operating cash flow. The QA module retrieves values from the normalized schema, enabling rapid validation without scanning raw worksheets.

Quality and Audit Signals Auditability is a primary design goal. We define a unified status ontology `OK`, `MISSING`, `PARSE_ERROR`, or `NOT_APPLICABLE` where `NOT_APPLICABLE` is semantic rather than tied to literal wording in filings. The US/JP adapters predominantly instantiate OK/MISSING, while CN additionally activates `PARSE_ERROR` and `NOT_APPLICABLE` via PDF-oriented status tracing. Summary indicators report missing-field counts for quick completeness assessment. The interface also supports downloading structured workbooks for offline verification.

Implementation The demo is a no-database web service that reads structured Excel artifacts generated by the batch pipeline, ensuring consistency with experimental results. The backend is implemented in Python, and the frontend is a lightweight static page with API endpoints for market listing, company selection, statement retrieval, QA, and workbook download. The demo can be launched locally with a single command and accessed remotely via SSH port forwarding, enabling reproducible presentations.

Demo Flow A typical session proceeds as follows: (1) select market and company, (2) inspect IS/BS/CF tabs and metadata, (3) issue QA prompts for core metrics, (4) download the workbook for manual cross-checking. This workflow emphasizes usability, explicit quality signals, and direct support for human verification.

6 Conclusion and Future Work

We presented FinReporting, an agentic workflow for cross-jurisdiction financial reporting. By canonicalizing heterogeneous filings into a unified IS/BS/CF representation and decomposing processing into auditable stages with explicit verification signals, FinReporting addresses key frictions in global financial analysis. Unlike free-form LLM extraction pipelines, FinReporting uses the LLM as a constrained verifier, repairing outputs only when sufficient evidence is available and deferring uncertain cases to human review. This design enables transparent cross-market reporting, surfaces aggregation and mapping anomalies, and provides a scalable, auditable infrastructure that reduces reliance on manual data curation and mitigates hidden structural errors downstream.

Future work will expand jurisdiction coverage and enrich the canonical ontology to capture long-tail taxonomy variations, extend verification to footnotes and segment disclosures, and improve robustness across statements, periods, and noisy PDFs.

Limitations

We acknowledge limitations of FinReporting. First, the current implementation focuses on annual filings from three jurisdictions (US, JP, and CN) and a fixed set of core IS/BS/CF targets. Extending to additional markets, reporting standards, and finer-grained line items remains future work.

Second, although XBRL-native jurisdictions provide structured inputs, PDF-centric environments (e.g., CN filings) introduce layout variability, fragmented tables, OCR noise, and issuer-specific conventions that reduce extraction reliability. In such cases, the system may defer to NEED_REVIEW.

Third, canonical alignment relies on a predefined financial ontology, which may not fully capture market-specific nuances, long-tail taxonomy variations, or firm-specific disclosure idiosyncrasies.

Finally, FinReporting is an auditable reporting assistant, not a fully autonomous system; human verification is necessary for high-stakes use cases.

Ethical Considerations

FinReporting operates in a high-stakes financial domain where mislocalized or incorrectly extracted values could mislead downstream analysis. To mitigate such risks, the system explicitly surfaces uncertainty through status labels, anomaly logs, and review flags, and constrains the LLM to bounded verification actions rather than free-form generation.

The demo system is intended for structured reporting assistance and inspection. It should not be used as the sole basis for investment or regulatory decisions without independent validation against original filings.

We also note that cross-jurisdiction localization necessarily involves modeling choices through a canonical ontology. While designed to preserve semantic faithfulness, any abstraction layer may introduce interpretative bias, especially in edge cases involving jurisdiction-specific accounting conventions. Users should therefore treat localized outputs as aligned representations rather than authoritative accounting restatements. Additionally, users are encouraged to review provenance metadata and cross-check anomalies to ensure transparency, traceability, and contextual correctness in interpretations.

Data License We use publicly available regulatory disclosures, including US filings from SEC EDGAR, Japanese filings from EDINET, and Chinese annual report PDFs from official public disclosure platforms. These documents are subject to the terms and conditions of their respective providers.

Our released artifacts (e.g., code, canonical schemas, evaluation templates, structured outputs, and audit logs) do not include raw proprietary documents and can be redistributed independently of the original filings. Users are responsible for obtaining source filings directly from official regulatory portals in compliance with applicable data usage policies.

References

Toyin D. Aguda, Suchetha Siddagangappa, Elena Kochkina, Simerjot Kaur, Dongsheng Wang, and Charese Smiley. 2024. Large language models as financial data annotators: A study on effectiveness and efficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, LREC-COLING '24, pages 10124–10145, Torino, Italia. ELRA and ICCL.

Jian Chen, Peilin Zhou, Yining Hua, Loh Xin, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024. FinTextQA: A dataset for long-form financial question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, ACL '24, pages 6025–6047, Bangkok, Thailand. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP '21, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, EMNLP '22, pages 6279–6292, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

European Securities and Markets Authority. 2025. ESEF XBRL taxonomy 2024 documentation. https://www.esma.europa.eu/sites/default/files/2025-01/esef_taxonomy_2024_documentation.pdf. Accessed: 2026-02-26.

Shijie Han, Haoqiang Kang, Bo Jin, Xiao-Yang Liu, and Steve Y. Yang. 2024. XBRL agent: Leveraging large language models for financial report analysis. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, pages 856–864, Brooklyn, NY, USA. Association for Computing Machinery.

Yusheng Huang, Ning Hu, Kunping Li, Nan Wang, and Zhouhan Lin. 2024. Extracting financial events from raw texts via matrix chunking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, LREC-COLING '24, pages 7035–7044, Torino, Italia. ELRA and ICCL.

IFRS Foundation. 2024. IFRS accounting taxonomy 2024. https://www.ifrs.org/issued-standards/ifrs-taxonomy/ifrs-accounting-taxonomy-2024/. Published: 2024-03-27. Accessed: 2026-02-26.

Fengxian Ji, Jingpu Yang, Zirui Song, Lang Gao, Junhong Liang, Zhenhao Chen, Jinghui Zhang, and Xiuying Chen. 2026. ServImage: An image generation and editing benchmark from real-world commercial imaging services. *ArXiv preprint*, arXiv:2604.24023.

Fengxian Ji, Jingpu Yang, Zirui Song, Yuanxi Wang, Zhexuan Cui, Yuke Li, Qian Jiang, Miao Fang, and Xiuying Chen. 2025. FineState-Bench: A comprehensive benchmark for fine-grained state control in GUI agents. *ArXiv preprint*, arXiv:2508.09241.

Subhendu Khatuya. 2024. Parameter efficient instruction tuning of LLMs for financial applications. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, pages 8494–8495, Jeju, South Korea. International Joint Conferences on Artificial Intelligence Organization. Doctoral Consortium.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL-HLT '09, pages 272–280, Boulder, Colorado. Association for Computational Linguistics.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65.

Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Vincent Jim Zhang, Yuqing Guo, Jeff Zhao, Huan He, Yi Han, Yun Feng, Yuechen Jiang, Yupeng Cao, Haohang Li, Yangyang Yu, Xiaoyu Wang, Penglei Gao, and 28 others. 2025. MultiFinBen: Benchmarking large language models for multilingual and multimodal financial application. *ArXiv preprint*, arXiv:2506.14028.

Lingfei Qian, Xueqing Peng, Yan Wang, Vincent Jim Zhang, Huan He, Hanley Smith, Yi Han, Yueru He, Haohang Li, Yupeng Cao, Yangyang Yu, Alejandro Lopez-Lira, Peng Lu, Jian-Yun Nie, Guojun Xiong, Jimin Huang, and Sophia Ananiadou. 2025a. When agents trade: Live multi-market trading benchmark for LLM agents. *ArXiv preprint*, arXiv:2510.11695.

Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Han Yi, Yilun Zhao, Jimin Huang, Qianqian Xie, and Jian-Yun Nie. 2025b. Fino1: On the transferability of reasoning-enhanced LLMs and reinforcement learning to finance. *ArXiv preprint*, arXiv:2502.08127.

Ruoqi Shu, Xuhui Wang, Isaac Wang, Yanming Mai, and Bo Wan. 2025. LAVA: Logic-aware validation and augmentation framework for large-scale financial document auditing. In *Proceedings of the 10th Workshop on Financial Technology and Natural Language Processing*, FinNLP '25, pages 75–92, Suzhou, China. Association for Computational Linguistics.

U.S. Securities and Exchange Commission. 2025a. Financial statement and notes data sets. `https://www.sec.gov/data-research/sec-markets-data/financial-statement-notes-data-sets`.

U.S. Securities and Exchange Commission. 2025b. Financial statement data sets. `https://www.sec.gov/data-research/sec-markets-data/financial-statement-data-sets`.

Xinlin Wang and Mats Brorsson. 2025. Can large language model analyze financial statements well? In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing, the 6th Financial Narrative Processing, and the 1st Workshop on Large Language Models for Finance and Legal*, FinNLP-FNP-LLMFinLegal '25, pages 196–206, Abu Dhabi, UAE. Association for Computational Linguistics.

Yan Wang, Lingfei Qian, Xueqing Peng, Yang Ren, Keyi Wang, Yi Han, Dongji Feng, Fengran Mo, Shengyuan Lin, Qinchuan Zhang, Kaiwen He, Chenri Luo, Jianxing Chen, Junwei Wu, Chen Xu, Ziyang Xu, Jimin Huang, Guojun Xiong, Xiao-Yang Liu, and 2 others. 2025. FinTagging: Benchmarking LLMs for extracting and structuring financial information. *ArXiv preprint*, arXiv:2505.20650.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David S. Rosenberg, and Gideon Mann. 2023. BloombergGPT: A large language model for finance. *ArXiv preprint*, arXiv:2303.17564.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, and 15 others. 2024. FinBen: A holistic financial benchmark for large language models. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*, NeurIPS '24, Vancouver, BC, Canada.

Zhuohan Xie, Dhruv Sahnan, Debopriyo Banerjee, Georgi Georgiev, Rushil Thareja, Hachem Madmoun, Jinyan Su, Aaryamonvikram Singh, Yuxia Wang, Rui Xing, Fajri Koto, Haonan Li, Ivan Koychev, Tanmoy Chakraborty, Salem Lahlou, Veselin Stoyanov, and Preslav Nakov. 2026. FinChain: A symbolic benchmark for verifiable chain-of-thought financial reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, ACL '26, San Diego, California, USA.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-source financial large language models. *ArXiv preprint*, arXiv:2306.06031.

Yixi Zhou, Fan Zhang, Yu Chen, Haipeng Zhang, Preslav Nakov, and Zhuohan Xie. 2026. FinCARDS: Card-based analyst reranking for financial document question answering. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, California, USA.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '21, pages 3277–3287, Online. Association for Computational Linguistics.