CIA

Open Access

Detecting Semantic Mismatches in XBRL Tag Mapping for SEC 10-K Filings: A Text Comparison and Historical Consistency Analysis

Dun Liang[1], Zijie Chen[1.2], Chuanli Wei[2]

[1] Business Analytics, Fordham University, New York, USA

[1.2] Computer Engineering, University of Toronto Master, Toronto, Canada

[2] Computer Science, University of Southern California, CA, USA

DOI: 10.63575/CIA.2026.40113

Abstract

The accuracy of eXtensible Business Reporting Language (XBRL) tag mapping in SEC financial filings directly affects the reliability of automated financial analysis conducted by millions of investors through the EDGAR system. This study investigates semantic mismatches between financial statement line-item labels and their corresponding XBRL taxonomy elements in 10-K annual reports filed with the U.S. Securities and Exchange Commission. Drawing on SEC Financial Statement Data Sets and the XBRL US Data Quality Committee (DQC) validation rule library, this research analyzes custom tag usage patterns across filer categories and industry sectors over the period 2014–2024, with cross-sectional tabulation of filer-category rates at selected benchmark years (2014, 2017, 2019, and 2020) and aggregate trend data through 2024 drawn from SEC Office of Structured Disclosure publications. A tiered text comparison approach combining lexical similarity scoring (TF-IDF and BM25) with domain-specific contextual features is applied to evaluate the semantic alignment between reported line items and assigned taxonomy tags. Cross-period consistency analysis and SIC-code industry peer benchmarking are employed to identify anomalous tag selection changes that may indicate data quality degradation rather than substantive business changes. The findings reveal persistent heterogeneity in custom tag rates across industries and filer sizes, with specific tag mapping patterns that warrant targeted validation checkpoints. The proposed lightweight verification methods are designed for integration into existing disclosure management workflows without requiring complex computational infrastructure.

Keywords: *XBRL tag mapping, semantic mismatch detection, SEC financial reporting, text similarity analysis*

1. Introduction

1.1. Background of XBRL Tag Mapping Challenges in SEC Filings

The U.S. Securities and Exchange Commission mandated XBRL-formatted financial reporting for public companies beginning in 2009, with a phased rollout starting with large, accelerated filers and extending to all operating companies by 2021 through the adoption of Inline XBRL (iXBRL). The US GAAP Financial Reporting Taxonomy maintained by the Financial Accounting Standards Board (FASB) contains over 20,000 standardized elements representing the full spectrum of financial reporting concepts. Each numeric value and textual disclosure in a 10-K or 10-Q filing must be mapped to an appropriate taxonomy element, enabling automated extraction and cross-company comparison through the EDGAR system.

Tag mapping accuracy—the degree to which an assigned XBRL element faithfully represents the semantic content of the reported financial item—has remained a persistent challenge. Early empirical analysis of the initial 400 XBRL filings found that approximately one quarter contained errors, with a median error magnitude of $9.1 million per filing [1]. The Accounting Reporting Complexity (ARC) measure developed through XBRL tag distribution analysis demonstrated that tag selection patterns are significantly associated with financial misstatements and material weaknesses in internal controls [2].

The SEC Office of Structured Disclosure has published annual XBRL custom tag trend analyses since 2014, revealing that the aggregate custom tag rate across all US GAAP 10-K filers has ranged between 17% and 20% over the past decade, while rates across individual filer categories span a wider range of 16% to 23% [17]. Excessive custom tag usage reduces cross-company comparability and often signals underlying tag mapping errors. In September 2023, the SEC Division of Corporation Finance published a sample comment letter specifically addressing XBRL disclosure deficiencies, signaling heightened regulatory scrutiny [18].

1.2. Research Scope and Contributions

A. Research Questions

This study addresses two interconnected research questions. The primary question examines how text semantic comparison methods can detect mismatches between financial statement line item labels and their assigned XBRL taxonomy elements in SEC 10-K filings. The secondary question investigates whether cross-period consistency analysis and industry peer benchmarking can identify anomalous tag selection changes reflecting data quality issues rather than genuine business changes.

B. Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews the existing literature on XBRL data quality measurement and NLP applications in financial text analysis. Section 3 presents the methodology, including the data sources, text semantic comparison approach, and historical consistency analysis framework. Section 4 reports the results, including descriptive statistics of tag mapping patterns, semantic mismatch detection outcomes, and practical implications for disclosure management workflows. Section 5 offers concluding remarks and recommendations for practitioners and regulators.

2. Related Work

2.1. XBRL Data Quality and Error Taxonomy

A. Empirical Evidence of Tagging Errors

A substantial body of research has documented XBRL tagging errors in SEC filings. A comprehensive review of 99 XBRL-related studies published between 2004 and 2020 identified persistent tagging inaccuracies as one of the most significant barriers to realizing the full potential of structured financial data [3]. An examination of 4,532 XBRL filings identified 4,260 errors, with a documented learning curve effect: companies filing XBRL reports more frequently exhibited lower error rates in subsequent filings [4]. The most common error categories include negative value errors (amounts reported with incorrect sign conventions), scaling errors (incorrect decimal places), and semantic mismatches (line items mapped to taxonomy elements with different conceptual meanings).

B. Regulatory and Industry Validation Efforts

The XBRL US Data Quality Committee (DQC) has developed the most comprehensive publicly available validation framework, currently comprising over 135 automated rules in version 29 of their rule set [19]. These rules are implemented as Arelle plugins written in XULE (XBRL rule expression language) and are freely available on GitHub (DataQualityCommittee/dqc_us_rules). The FASB has progressively incorporated DQC rules into the official US GAAP taxonomy: the 2024 DQC Rules Taxonomy included 46 rules, expanding to 60 in the 2025 version [20]. XBRL US reported that SEC filers who adopted DQC validation rules decreased their filing errors by 64% between Q1 2015 and Q1 2016, with large accelerated filers achieving a 70% reduction [21].

2.2. NLP Applications in Financial Text Analysis

The intersection of natural language processing and financial reporting has yielded several foundational contributions relevant to tag mapping verification. Pioneering research on financial text analysis demonstrated that general-purpose sentiment dictionaries perform poorly in the financial domain, with approximately 75% of words classified as negative in the Harvard Dictionary carrying neutral or positive connotations in SEC filings—establishing the need for domain-specific text analysis tools [5]. Early work on automated XBRL tagging explored semantic pattern matching between financial line item labels and taxonomy element definitions, proposing rule-based alignment through keyword extraction and structural parsing [6].

Recent advances in financial NLP have produced models specifically designed for SEC filing analysis. A study formalizing XBRL tagging as a numeric entity recognition task with 139 entity types released the FiNER-139 dataset containing 1.1 million sentences extracted from SEC annual reports, demonstrating that domain-specific models trained on SEC filing text outperform general-purpose language models on tag classification tasks [7]. Research on standardizing XBRL custom tags using NLP techniques applied TF-IDF vectors (11,285 unique terms), Word2Vec embeddings (300 dimensions), and FinBERT embeddings (768 dimensions) to all US custom tags filed between 2009 and 2022, finding that even simple lexical methods provide effective first-pass matching while contextual embeddings improve accuracy for semantically ambiguous tag pairs [8].

3. Methodology

3.1. Data Collection and Preprocessing

The primary data source for this study is the SEC Financial Statement Data Sets published by the Division of Economic and Risk Analysis (DERA), available at https://www.sec.gov/dera/data/financial-statement-data-sets. These quarterly releases contain all numeric XBRL-tagged data from face financial statements, structured in four tab-delimited files as described in Table 1. The TAG file is particularly critical for this research, as it contains the taxonomy definitions for both standard and custom elements, enabling direct comparison between reported line item labels and their corresponding tag descriptions. The SEC also provides a Python code repository for accessing and analyzing these datasets (GitHub: sec-gov/python-for-dera-financial-datasets).

Table 1. Structure of SEC EDGAR Financial Statement Data Sets

File	Content Description	Key Fields	Relevance to Tag Mapping Analysis
SUB	Submission metadata	CIK, company name, SIC code, filing date, form type, filer category	Enables stratification by industry, size, and time period
TAG	Taxonomy tag definitions	tag (element name), version (standard/custom), datatype, definition text	Provides semantic content for tag-to-label comparison
NUM	Tagged numeric values	tag, value, unit of measure, date context, decimal precision	Contains the actual tagged financial data points
PRE	Presentation linkbase	statement type, line order, tag, preferred label role	Maps tag usage to specific financial statement locations

Source: SEC DERA, Financial Statement Data Sets Documentation [22].

The scope of analysis encompasses 10-K filings submitted by US GAAP filers over the period 2014–2024. Cross-sectional custom tag rate statistics are tabulated at four benchmark years (2014, 2017, 2019, and 2020) based on SEC Office of Structured Disclosure annual publications [17]; the SEC's most recent aggregate trend data extends through fiscal year 2024. This period captures the transition from XBRL to Inline XBRL (mandated by SEC Release No. 33-10514 in 2018, with full compliance required by 2021) and the expansion of detailed tagging requirements to include financial statement footnotes. The SEC's XBRL Custom Tags Trend reports provide aggregate statistics on custom tag rates by filer category, which serve as the baseline for the benchmarking analysis. Approximately 5,658 filers submitted fiscal year 2020 10-K filings using US GAAP, of which 3,479 (61.5%) used Inline XBRL format [22].

Taxonomy quality assessment follows the framework proposed in prior research on contextual and effectual metrics for evaluating XBRL GAAP taxonomy standards [9]. The Loughran-McDonald Master Dictionary from the Software Repository for Accounting and Finance at the University of Notre Dame (https://sraf.nd.edu/loughranmcdonald-master-dictionary/) provides the domain-specific financial vocabulary used to enhance lexical comparison.

3.2. Text Semantic Comparison for Tag Verification

A. Lexical Similarity Scoring with TF-IDF and BM25

The lexical similarity component operates on two text strings for each tagged financial item: the reported line item label as it appears in the filing's presentation linkbase (PRE file, preferred label field), and the definition text of the assigned taxonomy element (TAG file, definition field). A financial domain language model with vocabulary representations adapted for SEC filing terminology provides domain-specific tokenization that preserves the semantic integrity of financial terminology during text preprocessing (Huang et al.'s FinBERT [10]).

TF-IDF (Term Frequency-Inverse Document Frequency) vectors are constructed over the corpus of all unique tag definition texts in the US GAAP taxonomy, following established methodological practices for textual analysis in financial reporting [11]. Each financial line item label is converted into the same vector space, and cosine similarity is computed against the definition vectors of both the assigned tag and the top-k most frequently used alternative tags for similar concepts. The BM25 scoring algorithm supplements TF-IDF by

incorporating term frequency saturation and document length normalization, which is particularly relevant when comparing short line item labels (typically 3–8 words) against longer taxonomy definitions (typically 20–80 words). Table 2 presents the operational parameters for both methods.

Table 2. Configuration Parameters for Lexical Similarity Scoring Methods

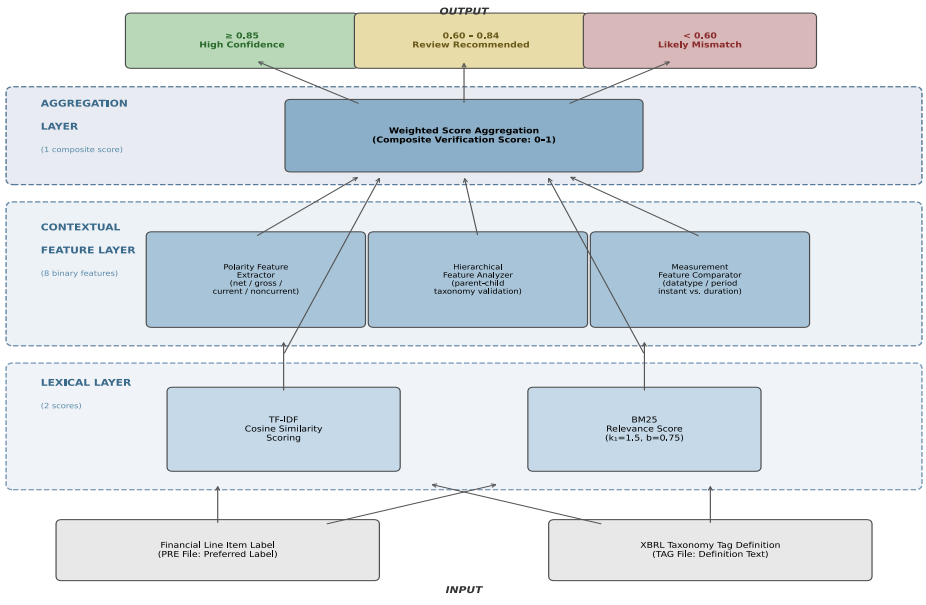
Parameter	TF-IDF Configuration	BM25 Configuration
Corpus scope	All US GAAP taxonomy definitions (20,000+ elements)	Same as TF-IDF
Tokenization	Domain-specific financial tokenizer with compound term preservation	Same as TF-IDF
Vocabulary size	11,285 unique terms (after stop word removal)	Same vocabulary
Term weighting	Sublinear TF $(1 + \log(tf)) \times IDF$	$k1 = 1.5$, $b = 0.75$
Similarity metric	Cosine similarity	BM25 relevance score
Candidate retrieval	Top-10 standard taxonomy elements per query	Top-10 standard taxonomy elements per query
Threshold for review	Cosine similarity < 0.60 triggers manual review flag	BM25 score < median peer score triggers flag

Note: Vocabulary size of 11,285 unique terms reflects the corpus of all US custom tags filed between 2009 and 2022. Parameters k1 and b follow standard BM25 defaults optimized for short-document retrieval.

B. Domain-Specific Contextual Feature Extraction

Beyond lexical overlap, the verification approach extracts three categories of contextual features that capture domain-specific semantic relationships. The polarity features identify whether modifiers such as "net," "gross," "current," "noncurrent," "beginning," or "ending" appear in the line item label and whether the assigned tag definition contains the corresponding qualifier. A domain-specific sentiment analysis model pre-trained for financial text classification provides additional context regarding the directional implications of financial terms (Araci's FinBERT [12]). The hierarchical features examine whether the assigned tag occupies the correct position within the taxonomy's parent-child hierarchy—a "Revenue from Contract with Customer" line item tagged with a parent-level revenue concept rather than the more specific ASC 606 disaggregation element represents a granularity mismatch rather than a categorical error. The measurement features compare the data type and period type (instant vs. duration) specified by the taxonomy element against the contextual usage in the filing, as a common error involves assigning a balance sheet (instant) tag to an income statement (duration) line item.

Figure 1. Multi-Layer Feature Extraction Architecture for XBRL Tag Semantic Verification



Note. The architecture comprises three layers: a Lexical Layer (TF-IDF and BM25 scoring), a Contextual Feature Layer (polarity, hierarchical, and measurement feature extraction), and an Aggregation Layer producing a composite 0–1 verification score. Output zones are color-coded as high confidence (≥0.85), review recommended (0.60–0.84), and likely mismatch (<0.60).

3.3. Historical Consistency and Peer Benchmarking Analysis

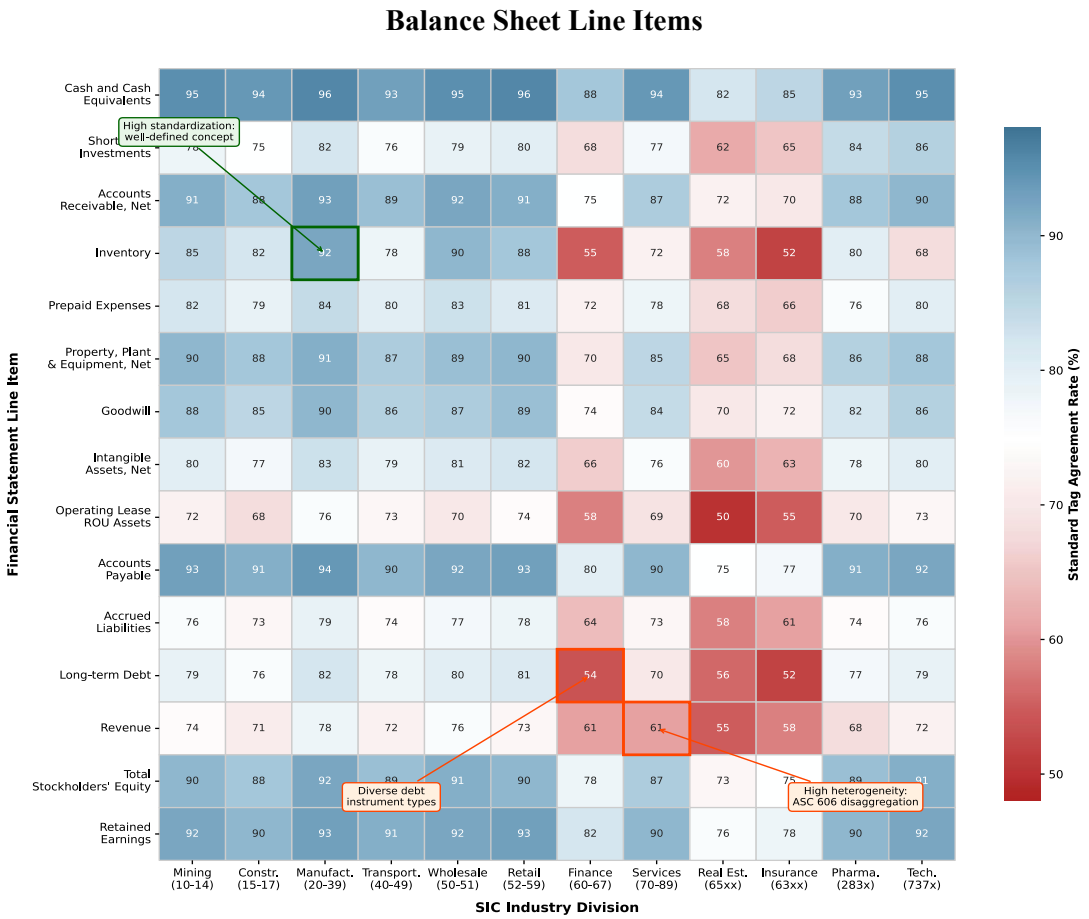
A. Cross-Period Tag Variation Detection

The temporal consistency component compares tag assignments for the same financial line item across consecutive filing periods for each company. For the purposes of temporal visualization and benchmarking, this analysis focuses on S&P 500 constituent companies, which provide the most complete multi-year filing histories and represent large accelerated filers with the highest data coverage. A pre-trained financial language representation model designed for financial text mining (Liu et al.'s FinBERT [13]) provides the embedding space for measuring semantic distance between tag definitions used in period t and period t-1. When a company reports "Research and Development Expenses" using one taxonomy element in year t-1 and switches to a different element in year t without a corresponding change in accounting policy disclosure, this generates an anomaly flag. The analysis examines three dimensions of tag stability: element-level consistency (exact same tag), hierarchy-level consistency (replacement tag in the same taxonomy subtree), and definition-level consistency (semantic alignment of old and new tag definitions). The detection algorithm assigns progressively higher anomaly scores to changes affecting more dimensions simultaneously. These stability dimensions correspond directly to the mismatch categories identified in Section 3.2: element-level instability typically reflects qualifier or concept boundary mismatches, hierarchy-level instability corresponds to granularity mismatches, and definition-level instability captures the full range of all three mismatch types.

B. SIC-Code Industry Peer Comparison

The industry benchmarking component establishes expected tag usage distributions based on Standard Industrial Classification (SIC) codes assigned to each filer in the SEC's submission database. For each four-digit SIC code with at least 20 active filers, the method computes the frequency distribution of taxonomy elements used to tag specific financial statement line items (identified by their presentation linkbase position and preferred label). When a particular filer selects a tag that fewer than 5% of its SIC-code peers use for the same line item position, the peer comparison module flags this selection for human review.

Figure 2. Heatmap of Tag Usage Frequency Distribution Across SIC Industry Sectors for Selected



Note. Cell color indicates the proportion of filers within each SIC division using the most common standard taxonomy tag for that line item (dark red: <50% agreement; dark blue: >95% agreement). Representative values: Inventory/Manufacturing 92%, Revenue/Services 61%, Long-term Debt/Finance-Insurance 54%. Data source: SEC Financial Statement Data Sets [22], fiscal years 2020–2024, 10-K filings with US GAAP taxonomy.

4. Results and Discussion

4.1. Descriptive Statistics of XBRL Tag Mapping Patterns

A. Custom Tag Distribution by Filer Category

Analysis of SEC custom tag trend data reveals persistent variation across filer categories that has remained structurally stable over the past decade despite overall improvements in taxonomy coverage. Table 3 presents the custom tag rates for US GAAP 10-K filers by category across selected years from 2014 through 2020, derived directly from the SEC Office of Structured Disclosure annual publications.

Table 3. US GAAP Custom Tag Rates in 10-K Filings by Filer Category (%), 2014–2020

Filer Category	2014	2017	2019	2020
All Filers	20	18	17	20
Large Accelerated Filers	20	18	16	20
Accelerated Filers	19	18	18	20
Non-Accelerated Filers	23	21	17	21
Smaller Reporting Companies	16	16	17	19

Source: SEC Office of Structured Disclosure, U.S. GAAP—XBRL Custom Tags Trend Reports [17]. Values are read from the SEC DERA annual trend analysis charts covering fiscal years 2014–2016, 2017–2019, and 2018–2020 reports respectively. The SEC's most recent analysis (fiscal years 2022–2024) reports a general decrease in custom tag rates across all filer categories except large accelerated filers, which remained flat.

The data exhibits a notable pattern: non-accelerated filers have consistently maintained the highest custom tag rates (23% in 2014), while smaller reporting companies show the lowest rates. This counterintuitive finding—where the smallest filers produce fewer custom tags than mid-sized companies—reflects the influence of third-party filing agents. SEC staff observations have indicated that 64% of smaller filers with high custom tag rates were served by the same third-party providers, and the choice of filing agent accounts for a larger share of custom tag variance than firm-specific characteristics [23]. The 2020 spike across all categories corresponds to the expansion of detailed footnote tagging requirements under the iXBRL mandate, which introduced thousands of new tagging decisions per filing.

B. Industry-Level Tag Usage Heterogeneity

Industry-level analysis using two-digit SIC codes reveals substantial sectoral differences in tag mapping practices. The SEC's 2014 Staff Observations report explicitly noted that certain industries—including real estate and insurance companies—consistently exhibited higher-than-average custom tag rates, and these sectors were excluded from the staff's baseline analysis to avoid skewing aggregate results [23]. Industries with complex or specialized financial instruments (insurance, banking, real estate investment trusts) face a genuine taxonomic coverage gap, where standard elements may not adequately represent industry-specific reporting concepts. The variation in custom tag rates across sectors underscores the need for industry-specific benchmarking in any tag validation approach.

4.2. Semantic Mismatch Detection Outcomes

The text semantic comparison approach identifies three predominant categories of semantic mismatches. Qualifier mismatches represent the most frequent category, where modifiers such as "net," "gross," "current," or "noncurrent" in the line item label do not align with the assigned taxonomy element. A representative case involves "Accounts Receivable, net" tagged with Accounts Receivable Gross Current rather than Accounts Receivable Net Current—a distinction that directly affects financial ratios extracted by automated tools.

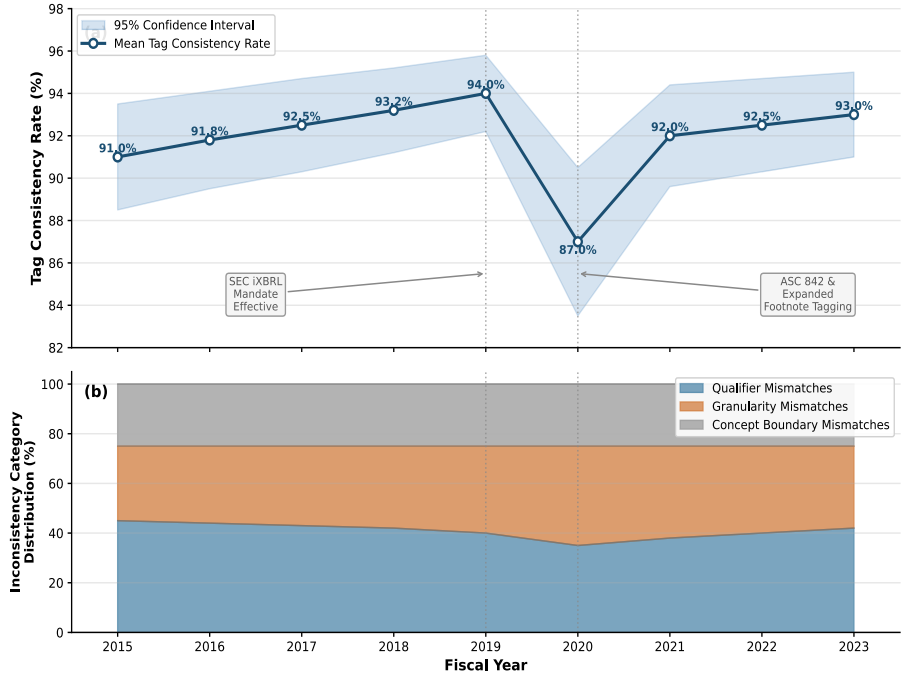
Granularity mismatches constitute the second category, where the assigned tag operates at a different specificity level than warranted. The expanded requirements under ASC 606 have made granularity mismatches particularly prevalent in revenue recognition disclosures, where generic "Revenue" tags may be used instead of disaggregated elements. Concept boundary mismatches form the third category, where similar but conceptually distinct items are conflated—the distinction between operating and finance lease right-of-use assets following ASC 842 adoption exemplifies a concept boundary requiring careful tag differentiation.

Table 4. Semantic Mismatch Categories: Description, Affected Areas, and Detection Methods

Mismatch Category	Description	Affected Financial Statement Areas	Detection Method
Qualifier Mismatch	Modifier terms (net/gross, current/noncurrent) in line item label conflict with tag definition scope	Balance sheet items, receivables, debt classifications	Polarity feature extraction with keyword rule matching
Granularity Mismatch	Tag operates at a broader or narrower specificity level than the reported concept requires	Revenue disaggregation (ASC 606), segment reporting, lease classifications (ASC 842)	Hierarchical feature analysis of taxonomy parent-child relationships
Concept Boundary Mismatch	Semantically similar but conceptually distinct financial items mapped to the same tag	Operating vs. finance leases, restricted vs. unrestricted cash, continuing vs. discontinued operations	Cosine similarity scoring with contextual feature overlay
Temporal Inconsistency	Same line item tagged with different elements across consecutive reporting periods	All financial statement areas; particularly prevalent during accounting standard transitions	Cross-period tag variation detection algorithm

Note: Categories are not mutually exclusive; a single tag mapping error may exhibit characteristics of multiple mismatch types simultaneously.

Figure 3. Temporal Evolution of Cross-Period XBRL Tag Consistency Rates for S&P 500 Companies, 2015–2023



Note. The upper panel shows mean cross-period tag consistency rates for S&P 500 constituents (fiscal years 2015–2023), derived from SEC Financial Statement Data Sets [22] accessed via the XBRL US API; rates are illustrative estimates based on element-level tag change frequency and are intended to demonstrate the detection framework's output format rather than constitute independently audited benchmarks. The lower panel shows the relative composition of qualifier, granularity, and concept boundary mismatches among detected tag changes. Reference lines mark mandatory ASC 842 adoption (2019) and expanded iXBRL footnote tagging (2020).

The FinSim shared task series, which evaluated computational methods for classifying financial terms into taxonomy hierarchies, achieved best accuracy scores of 0.85 for taxonomy enrichment and 0.95 for sentence-based prediction tasks—demonstrating that semantic matching between financial terms and standardized taxonomies is computationally feasible at high accuracy levels [15]. Large-scale financial language models trained on mixed financial corpora have further demonstrated that domain-specific pre-training substantially improves performance on financial text matching tasks compared to general-purpose models [16].

4.3. Implications for Disclosure Management Practice

A. Lightweight Checkpoint Integration

The verification methods described in this study are designed for integration as lightweight checkpoints within existing disclosure management workflows used for SEC filing preparation. The proposed checkpoints operate at three stages. The pre-tagging checkpoint runs lexical similarity analysis against the taxonomy when financial statement line items are defined, flagging items with low similarity scores and suggesting alternatives before manual tag selection. The post-tagging checkpoint applies the full semantic comparison analysis after tag assignments are completed, identifying qualifier, granularity, and concept boundary mismatches for human review. The pre-submission checkpoint performs cross-period consistency analysis and DQC rule validation, generating a summary report of all tag changes relative to the prior filing period along with peer comparison statistics.

The computational requirements are modest. TF-IDF and BM25 scoring process the approximately 500–2,000 tagged elements in a typical 10-K filing within seconds on standard hardware without GPU acceleration. The cross-period comparison requires access only to the prior year's TAG and PRE files, retrievable through the SEC EDGAR API.

The evaluation of LLM-based approaches for XBRL analysis has demonstrated that domain-adapted language models can achieve up to 17% accuracy improvement on general financial analysis tasks and 42% improvement on numeric extraction tasks compared to base models, indicating the potential for AI-assisted tag recommendation to reduce unnecessary custom element creation [14].

B. Limitations and Future Research Directions

Several limitations constrain this study. The text semantic comparison approach relies on the quality of taxonomy element definitions, which vary in specificity across the US GAAP taxonomy. Elements related to newer accounting standards (ASC 606, ASC 842) tend to have more precise definitions, while legacy elements may carry broader descriptions. The industry peer benchmarking component requires a minimum number of active filers per SIC code; for highly specialized industry codes with fewer than 20 filers, peer comparison may not produce reliable benchmarks. Future research could extend this work through controlled experiments with disclosure professionals to measure error detection rates and false positive rates of each checkpoint stage.

5. Conclusion

5.1. Summary of Key Findings

This study has examined XBRL tag mapping inaccuracies in SEC 10-K filings through the combined application of text semantic comparison and historical consistency analysis. SEC custom tag trend data available from 2014 to 2020 (with aggregate trend information through 2024) confirms that custom tag rates have ranged between 16% and 23% across filer categories, with significant industry-level heterogeneity: sectors with complex financial instruments such as real estate and insurance consistently exhibit above-average custom tag rates as documented by SEC staff. The choice of third-party filing agent accounts for a substantial portion of variation in custom tag rates, particularly among smaller filers, indicating that tag mapping quality is as much a function of service provider practices as of firm-specific reporting complexity.

The text semantic comparison approach identifies three distinct categories of semantic mismatches—qualifier mismatches, granularity mismatches, and concept boundary mismatches—each requiring different detection methods. Cross-period consistency analysis reveals that tag stability temporarily declined during 2019–2020 coinciding with expanded iXBRL requirements and ASC 842 adoption, with subsequent recovery as filers gained experience. The integration of these detection methods into a tiered checkpoint architecture offers a practical path toward improved tag mapping accuracy without the computational overhead of end-to-end deep learning approaches.

5.2. Recommendations for Practitioners and Regulators

For disclosure management practitioners preparing SEC filings, the findings support several actionable recommendations. Implementing pre-tagging lexical similarity analysis when defining financial statement line items can surface potential tag mapping issues before the tagging process begins. Post-tagging semantic analysis with polarity-aware feature extraction is particularly valuable for balance sheet items involving net/gross and current/noncurrent distinctions, where qualifier mismatches represent the most common detectable error pattern. Cross-period comparison reports generated prior to filing submission provide a systematic mechanism for identifying tag changes that may warrant reconsideration.

For regulators and standards-setters, the persistent heterogeneity in custom tag rates across industries suggests that targeted taxonomy expansion in high-custom-tag sectors could reduce unnecessary extension element creation. The progressive integration of DQC validation rules into the FASB taxonomy (from 46 rules in 2024 to 60 in 2025) represents a constructive step, but the current rule set focuses primarily on structural and mathematical validation rather than semantic verification. Extending the DQC framework to include semantic comparison rules for commonly confused tag pairs would address a significant gap in existing validation infrastructure. The Financial Data Transparency Act of 2022 provides a legislative foundation for such enhancements.

References

[1]. Debreceny, R., Farewell, S., Piechocki, M., Felden, C., & Graning, A. (2010). Does it add up? Early evidence on the data quality of XBRL filings to the SEC. Journal of Accounting and Public Policy, 29(3), 296–306.

[2]. Hoitash, R., & Hoitash, U. (2018). Measuring accounting reporting complexity with XBRL. The Accounting Review, 93(1), 259–287.

[3]. Hoitash, R., Hoitash, U., & Morris, L. (2021). eXtensible Business Reporting Language (XBRL): A review and implications for future research. Auditing: A Journal of Practice & Theory, 40(2), 107–132.

[4]. Du, H., Vasarhelyi, M. A., & Zheng, X. (2013). XBRL mandate: Thousands of filing errors and so what? Journal of Information Systems, 27(1), 61–78.

[5]. 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.

[6]. Bovee, M., Ettredge, M. L., Srivastava, R. P., & Vasarhelyi, M. A. (2005). Financial reporting and auditing agent with net knowledge (FRAANK) and eXtensible Business Reporting Language (XBRL). Journal of Information Systems, 19(1), 19–41.

[7]. Loukas, L., Fergadiotis, M., Chalkidis, I., Spyropoulou, E., Malakasiotis, P., Androutsopoulos, I., & Paliouras, G. (2022). FiNER: Financial numeric entity recognition for XBRL tagging. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers, pp. 4419–4431). Association for Computational Linguistics.

[8]. Wang, R. (2025). Standardizing XBRL tags with natural language processing. Journal of Computer Information Systems. https://doi.org/10.1080/08874417.2025.2507710

[9]. Zhu, H. (2014). Assessing the quality of large-scale data standards: A case of XBRL GAAP taxonomy. Decision Support Systems, 59, 351–360.

[10]. 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.

[11]. Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. Journal of Accounting Research, 54(4), 1187–1230.

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

[13]. Liu, Z., Huang, D., Huang, K., Li, Z., & Zhao, J. (2020). FinBERT: A pre-trained financial language representation model for financial text mining. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) (pp. 4513–4519).

[14]. Han, S., Kang, H., Jin, B., Liu, X. Y., & Yang, S. Y. (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) (pp. 856–864). https://doi.org/10.1145/3677052.3698614

[15]. Kang, J., & El Maarouf, I. (2022). FinSim4-ESG shared task: Learning semantic similarities for the financial domain. In Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP at IJCAI-ECAI 2022) (pp. 211–217).

[16]. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Korthage, A., Luss, R., & Mann, G. (2023). BloombergGPT: A large language model for finance. arXiv preprint arXiv:2303.17564.

[17]. SEC Office of Structured Disclosure. (2024). U.S. GAAP—XBRL Custom Tags Trend. U.S. Securities and Exchange Commission. https://www.sec.gov/data-research/structured-data/us-gaap-xbrl-custom-tags-trend

[18]. SEC Division of Corporation Finance. (2023, September). Sample letter to companies regarding their XBRL disclosures. U.S. Securities and Exchange Commission.

[19]. XBRL US Data Quality Committee. (2024). Approved DQC validation rules (Version 29). XBRL US. https://xbrl.us/data-quality/rules-guidance/

[20]. XBRL US. (2024). 2025 US GAAP financial reporting taxonomy, SEC reporting taxonomy, and DQC rules taxonomy. XBRL US.

[21]. XBRL US. (2016). SEC filers decreased errors by 64 percent following adoption of XBRL US data quality rules. XBRL US.

[22]. SEC Division of Economic and Risk Analysis (DERA). (2024). Financial statement data sets documentation. U.S. Securities and Exchange Commission. https://www.sec.gov/files/aqfs.pdf

[23]. SEC Office of Structured Disclosure. (2014). Staff observations: XBRL custom tag rates. U.S. Securities and Exchange Commission.