

FinTagging: An LLM-ready Benchmark for Extracting and Structuring Financial Information

Yan Wang
The Fin AI
USA

Yang Ren
The Fin AI
USA

Lingfei Qian
The Fin AI
USA

Xueqing Peng
The Fin AI
USA

Keyi Wang
Columbia University
USA

Yi Han
Georgia Institute of Technology
USA

Dongji Feng
Gustavus Adolphus College
St.Pete, Minnesota, USA

Xiao-Yang Liu
Columbia University
USA

Jimin Huang
The Fin AI
USA

Qianqian Xie
The Fin AI
USA
xqq.sincere@gmail.com

Abstract

We introduce FINTAGGING, the first full-scope, table-aware XBRL benchmark designed to evaluate the structured information extraction and semantic alignment capabilities of large language models (LLMs) in the context of XBRL-based financial reporting. Unlike prior benchmarks that oversimplify XBRL tagging as flat multi-class classification and focus solely on narrative text, FINTAGGING decomposes the XBRL tagging problem into two subtasks: FinNI for financial entity extraction and FinCL for taxonomy-driven concept alignment. It requires models to jointly extract facts and align them with the full 10k+ US-GAAP taxonomy across both unstructured text and structured tables, enabling realistic, fine-grained evaluation. We assess a diverse set of LLMs under zero-shot settings, systematically analyzing their performance on both subtasks and overall tagging accuracy. Our results reveal that, while LLMs demonstrate strong generalization in information extraction, they struggle with fine-grained concept alignment, particularly in disambiguating closely related taxonomy entries. These findings highlight the limitations of existing LLMs in fully automating XBRL tagging and underscore the need for improved semantic reasoning and schema-aware modeling to meet the demands of accurate financial disclosure. Code is available at our GitHub repository¹ and data is at our Hugging Face repository².

1 Introduction

Automated tagging is essential in financial reporting, converting structured content such as tables and text into machine-readable data by linking each number to its meaning. Globally, over 2 million

¹<https://github.com/The-FinAI/FinTagging>

²<https://huggingface.co/collections/TheFinAI/fintagging-68270132372c6608ac069bef>

companies publish financial reports disclosing earnings, expenses, and liabilities, which provide essential information for investors, regulators, and analysts. However, inconsistent terminology poses challenges for reliable interpretation. To address this, the eXtensible Business Reporting Language (XBRL) was introduced in 1999 as a global standard for machine-readable financial data [20]. By applying structured tags, XBRL transforms fragmented disclosures into a consistent format, enabling seamless data exchange, cross-entity comparisons, and large-scale analysis. Yet, as shown in Figure 1, accurately tagging over 2,000 numerical facts per report to more than 10,000 standardized concepts remains labor-intensive and error-prone. In 2023 alone, over 6,500 tagging errors were identified across 33,000 SEC filings³.

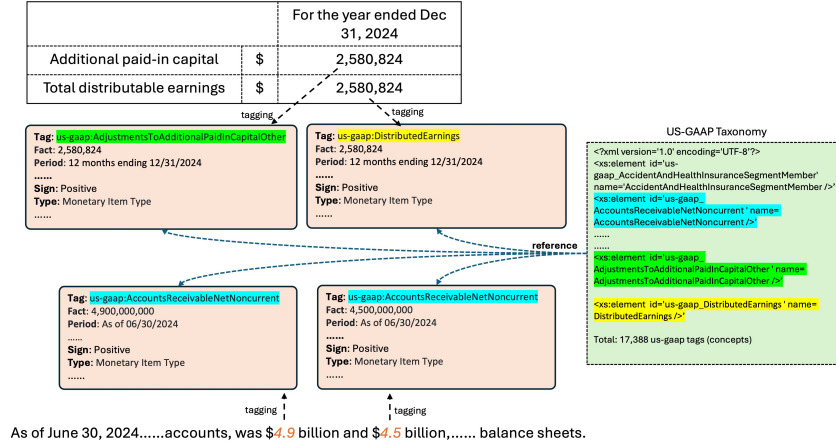


Figure 1: An example of Financial Tagging.

Existing approaches to automated XBRL tagging rely on task-specific models and pretrained language models (PLMs) [14, 23], but they require costly annotations, struggle to generalize to new filings, and perform poorly as taxonomies scale. In contrast, recent large language models (LLMs) show strong zero- and few-shot reasoning without task-specific supervision [31, 3, 29, 18], yet their application to XBRL tagging remains underexplored due to two major limitations (Table 1). First, **task framing is oversimplified**: prior benchmarks like FiNER [14] and FNXL [23] treat tagging as extreme multi-class classification over flat label sets, offering little context for disambiguating fine-grained financial facts. These datasets typically cover only 1k+ concepts, leaving most of the 10k+ US-GAAP taxonomy untested. As shown in Table 8, SOTA LLMs such as DeepSeek-V3 [13] and GPT-4o [11] achieve 0.0 precision, recall, and F1 under this setting. Second, **structured data is ignored**: existing datasets exclude tables, despite their central role in financial reporting. Many key facts appear in structured formats, as shown in recent financial QA benchmarks [4, 35, 16], making this omission a key gap between benchmarks and real-world tagging needs.

Table 1: Detailed comparison of financial NLP benchmarks across task types, sources, and structural capabilities. “QA” means the question-answering task. “Num. Reasoning” indicates the numerical reasoning. “Struct. IE” denotes the structured information extraction.

Benchmark	Scenario	Data Source	Task	Modality	#Entity Label	#Taxonomy Label	Num. Reasoning?	Struct. IE?	Concept Linking?
FinQA [4]	Decision making	SEC 10-K	QA	text/table	0	0	✓	✗	✗
ConvFinQA [5]	Decision making	SEC 10-K	QA	text/table	0	0	✓	✗	✗
TAT-QA [35]	Financial analysis	Chinese financial reports	QA	text/table	0	0	✓	✗	✗
DocVQA [16]	Enterprise automation	Financial document images	QA	text/image	0	0	✓	✗	✗
FNER-ORD [22]	Financial tagging	Chinese financial disclosures	Classification	text	49	0	✗	✗	✗
FinRED [24]	Knowledge graph construction	English financial news	Classification	text	38	0	✗	✗	✗
FiNER [14]	XBRL tagging	SEC 10-K	Extreme classification	text	0	139	✓	✗	✗
FNXL [23]	XBRL tagging	SEC 10-K	Extreme classification	text	0	2,800	✓	✗	✗
FinTagging (ours)	Financial & XBRL tagging	SEC 10-K	IE + Alignment	text/table	5	17,388	✓	✓	✓

To address these issues, we present FINTAGGING, the first LLM-ready benchmark for full-scope, structure-aware XBRL tagging, which challenges models to perform end-to-end fact extraction and taxonomy alignment on corporate filings. As shown in Table 1, unlike prior work [22, 14, 23] that frames tagging as flat multi-class classification and focuses solely on narrative text, FINTAGGING requires models to jointly extract financial facts and align them with the US-GAAP taxonomy

³<https://xbrl.us/wp-content/uploads/2023/03/DQC-SECMeetingNotes-20240314.pdf>

across both unstructured text and structured tables. This is the first LLM-oriented formulation that scales to the full set of 10k+ taxonomy labels while avoiding the intractability of single-step classification, enabling more fine-grained evaluation of model performance. In collaboration with financial reporting experts, we decompose the task into two components: the first, **structured numerical information extraction**, assesses a model’s ability to identify key financial facts from both text and tables; the second, **fine-grained concept linking**, evaluates its ability to map each fact to the correct taxonomy concept among all semantically similar candidates. Unlike prior benchmarks limited to 1,000 frequent concepts, our setting covers the entire taxonomy, exposing the limitations of frequency-biased classification and demanding precise semantic reasoning [2]. We construct two evaluation sets: *FinNI-eval*⁴ for numerical fact extraction and *FinCL-eval*⁵ for concept linking, both derived from real XBRL submissions and annotated with gold-standard mappings. We also introduce a unified evaluation framework that jointly measures extraction accuracy and concept alignment, providing a rigorous zero-shot assessment of LLMs’ capabilities for practical, table-aware XBRL tagging.

We evaluate ten state-of-the-art LLMs under a zero-shot setting on three fronts: (1) end-to-end macro-F1 over the full FINTAGGING benchmark, (2) subtask-specific performance on FinNI and FinCL, and (3) ablation of our unified extraction-and-alignment evaluation framework. DeepSeek-V3 [13] and GPT-4o [11] achieve the highest macro-F1 scores, indicating that our benchmark design enables strong handling of both frequent and rare financial tags, an improvement over traditional PLM baselines evaluated under token-classification settings. In subtask analyses, these LLMs excel at numerical information extraction (FinNI) but continue to struggle with precise concept linking (FinCL), underscoring the remaining challenge of fine-grained semantic alignment. Crucially, our ablation study shows that, without the joint evaluation framework, even top LLMs produce invalid tagging outputs, highlighting the framework’s essential role in realistic assessment of XBRL tagging readiness.

We conclude our main contributions as follows: (1) We introduce FINTAGGING, the first LLM-ready benchmark for full-scope, table-aware XBRL tagging with a two-stage pipeline—so that modern LLMs can be probed in true zero-shot setting. For each stage we release a new, professionally annotated evaluation set: FinNI-eval and FinCL-eval. (2) We conduct an extensive evaluation of state-of-the-art LLMs in zero-shot settings, systematically analyzing their performance across information extraction and semantic alignment subtasks, as well as overall tagging accuracy. (3) Our experimental results reveal a substantial performance gap between LLMs and the task requirements, particularly in fine-grained semantic alignment. This highlights the limitations of current LLMs in complex financial applications and underscores the need for continued advancements in task-specific adaptation.

2 Related Work

XBRL Tagging Benchmarks Previous benchmarks have framed XBRL tagging as a large-scale classification problem, primarily focusing on entity recognition. FiNER [14] introduced a dataset of 1.1 million sentences from SEC filings, annotated with 139 XBRL entity types, addressing challenges like context-sensitive numeric entities and financial domain-specific expressions. Building on this, Sharma et al. [23] proposed the Financial Numeric Extreme Labelling (FNXL) dataset, expanding the label space to approximately 2,800 XBRL tags and employing extreme classification techniques to handle the increased scale. While these works emphasize the complexity of financial taxonomies, their flat extreme classification approach overlooks the reasoning and alignment needed for realistic XBRL tagging, limiting their suitability for evaluating large language models.

XBRL Tagging Methods To improve XBRL tagging accuracy, prior work has incorporated structured knowledge and domain-specific modeling. Saini et al. [21] proposed **GalaXC**, a graph neural network leveraging label hierarchies for improved classification. Ma et al. [15] enriched transformer models with taxonomy definitions to disambiguate semantically similar tags. Loukas et al. [14] introduced **SECBERT** and numeral-aware input transformations to enhance robustness. For extreme multi-label tagging, Sharma et al. [23] applied **AttentionXML** to focus on context-relevant segments. Wang [28] aligned custom tags with standard taxonomies via semantic similarity using TF-IDF, Word2Vec, and FinBERT. More recently, Han et al. [10] developed **XBRL-Agent**, an LLM-based

⁴<https://huggingface.co/datasets/TheFinAI/FinNI-eval>

⁵<https://huggingface.co/datasets/TheFinAI/FinCL-eval>

system for extracting insights from filings. These efforts highlight the shift toward taxonomy-aware and LLM-enhanced solutions for XBRL tagging.

Financial Evaluation Benchmarks Broader financial NLP benchmarks address information extraction, reasoning, and document understanding tasks. FiNER-ORD [22] and FinRED [24] focus on entity recognition and relation extraction, respectively. BizBench [12] assesses the quantitative-reasoning ability of LLMs for both business and finance. Pixiu [33] evaluates LLMs across classification, QA, and summarization, while FinQA [4] and ConvFinQA [5] focus on numerical reasoning over text and tables. TAT-QA [35] and DocVQA [16] target table-text joint understanding. These benchmarks highlight the importance of integrating structured and unstructured information, yet they overlook the taxonomy-driven fact alignment and do not support the structured output required for XBRL tagging.

3 FINTAGGING

3.1 Task Formulation

Formally, given a financial document D consisting of multiple textual contents and structured tables, a set of predefined value data types L specified by the XBRL specification, and a taxonomy database \mathcal{T} containing a set of financial semantic concepts, the XBRL tagging task is to identify all relevant numerical values in D and annotate each with a structured triplet $\{\text{Fact}, \text{Type}, \text{Tag}\}$. Here, Tag denotes the linked concept in \mathcal{T} (e.g. “us-gaap:CashAndDueFromBanks”, etc.), Fact represents the extracted numerical value as it appears in context (e.g., “2.5”, “10,000”, “two”, etc.), and Type specifies the corresponding data type in L (including *monetaryItemType*, *percentItemType*, *sharesItemType*, *perShareItem*, and *integerItemType*). Here, *monetaryItemType* refers to a financial amount, such as revenue or expenses, typically reported in currency units (e.g., USD). *percentItemType* represents a rate or ratio, expressed as a decimal (e.g., 25%). *sharesItemType* indicates the number of shares, such as stocks held or issued. *perShareItem* captures values reported on a per-share basis, like earnings per share. Lastly, *integerItemType* is used for whole number counts, such as the number of employees or transactions.

We formalize the task as a mapping:

$$f : (D, L, \mathcal{T}) \mapsto \{(\text{Fact}_i, \text{Type}_i, \text{Tag}_i)\}_{i=1}^n \quad (1)$$

where each triplet corresponds to a financial value mention extracted from D and semantically grounded in \mathcal{T} .

Inspired by information extraction and alignment works [32, 27, 17, 30] and after discussed with financial reporting specialists, we formulated FINTAGGING into two sub-tasks: **Financial Numeric Identification (FinNI)**, a multi-modal numerical information extraction task, detects numerical value in D and classifies each with its appropriate Type. **Financial Concept Linking (FinCL)**, a numerical entity normalization task, then associates each identified value with its most appropriate Tag in \mathcal{T} based on contextual and structural cues.

Financial Numeric Identification (FinNI). The first subtask of FINTAGGING focuses on identifying numerical values in a financial document and assigning each a coarse-grained value data type. This corresponds to detecting the Fact and Type components of each triplet $\{\text{Fact}, \text{Type}, \text{Tag}\}$ defined in the overall task.

We formalize this subtask as a mapping:

$$f_{\text{FinNI}} : (D = (S, T), L) \mapsto \{(e_i, l_i)\}_{i=1}^k \quad (2)$$

where S and T represent the textual and tabular components of the document, respectively, and L is the set of predefined value data types. Each e_i is a numerical entity extracted from either S or T , and $l_i \in L$ denotes its assigned data type.

Financial Concept Linking (FinCL) Building on the output of the FinNI subtask, the goal of FinCL is to semantically ground each identified numerical entity e by linking it to a concept \hat{c} in

a predefined financial taxonomy, which is the Tag component of each triplet $\{\text{Fact}, \text{Type}, \text{Tag}\}$ defined in the overall task. Formally, we define the mapping as:

$$f_{\text{FinCL}} : (e, l, C_e, \mathcal{T}) \mapsto \hat{c} \quad (3)$$

where e is a numerical entity identified in the document $D = (S, T)$, $l \in L$ is its predicted data type from FINNI, C_e denotes the contextual information surrounding e in D , and $\mathcal{T} = \{c_1, c_2, \dots, c_n\}$ is a financial taxonomy containing n uniquely defined and semantically grounded concepts. The model is required to assign a concept $\hat{c} \in \mathcal{T}$ that best reflects the meaning of e in its structural context.

3.2 Raw Data Collection

We collected 30 annual 10-K reports filed in 2024 by publicly listed companies from SEC ⁶, as summarized in Table 2. There are a total of 81,325 facts, of which 69,451 are linked to standard taxonomy tags and 11,874 to SEC extension tags. Furthermore, using the BeautifulSoup tool to parse these reports, we identified 76,835 narrative sentences (approximately 16 million characters) and 5,450 financial tables. The companies’ information is shown in Appendix C. These reports follow the XBRL standard. We include all sections that may contain XBRL-tagged content to ensure comprehensive coverage.

Table 2: The statistic of collected financial reports.

	Report type	Period	#Company	#Sentence	#Char	#Table	#Std Tag	#Custom Tag	#Total Tag
Information	10-k	2024-02-13 to 2025-02-13	30	76,835	16,595,283	5,450	69,451	11,874	81,325

3.3 FINTAGGING Data Annotation

To construct the FINTAGGING benchmark, we annotated a total of 3,354 sentences and 3,245 table sequences extracted from the narrative and tabular content of collected financial reports, as summarized in Table 3. Among them, 1,627 textual instances and 2,014 table sequences contain XBRL tagging information (positive), while the remaining 1,727 textual and 1,231 table instances do not (negative). On average, textual inputs contain 82.55 tokens, whereas table sequences are significantly longer, averaging 999.74 tokens. Based on this, we further construct two subtask-specific datasets, FinNI-eval and FinCL-eval, with the following steps. The statistics for these two datasets are shown in Table 4.

Table 3: The statistical information for the original dataset in our benchmark. Using the “cl100k_base” tokenizer to calculate tokens (\pm standard deviation).

Structure	Pos/Neg	#Instance	Avg. Tokens/S	Avg. Entities/S	Avg. Concepts/S	Total Entities	Unique Concepts
Sentence	Positive	1,627	82.55 ± 66.39	1.27 ± 1.84	1.27 ± 1.84	52,740	2,288
	Negative	1,727					
Table	Positive	2,014	999.74 ± 993.37	14.89 ± 23.50	14.89 ± 23.50		
	Negative	1,231					

For narrative sentences, we retain only textual segments that exceed 20 characters in length and contain numerical terms. For table cell sequences, we use specific labels such as <table>, <tr>, <td>, and <th> to preserve layout and structural information. Tables that contain only headers without any numerical values are considered invalid and thus excluded from our benchmark.

3.3.1 Curation of Numeric Entities and US-GAAP Concepts

Building on the initial filtering, we further refined the sentence and table set to retain only high-quality samples with gold-standard annotations of numerical entity types and their associated US-GAAP concepts. As shown in Table 3, we obtained 52,740 annotated numerical entities spanning 5 entity types, linked to a total of 2,288 unique US-GAAP concepts.

⁶<https://www.sec.gov/>

Specifically, we predefine the five of the most prevalent item types, including *monetaryItemType*, *percentItemType*, *sharesItemType*, *perShareItemType*, and *integerItemType*. We first excluded sentences and tables lacking US-GAAP labels. Then we conducted a statistical analysis of the entity types contained in the data, as summarized in Figure 2, to identify the most frequent types.

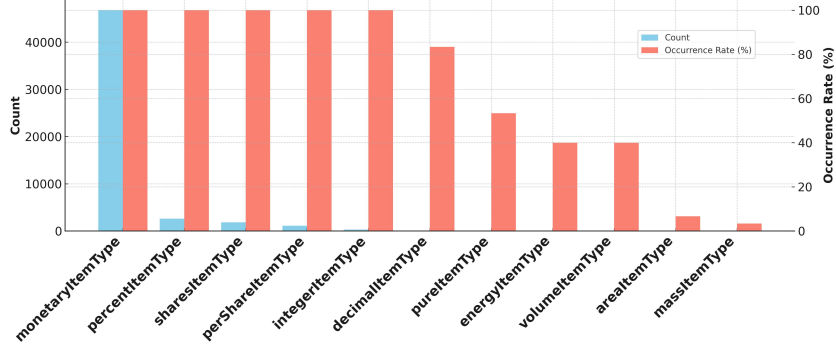


Figure 2: The statistic of numerical entity type.

From Figure 2, we identified 11 numerical entity types across the financial reports of 30 companies. The top five types, based on frequency and coverage, appear in all reports. Sentences or tables containing these types are labeled as positive instances; others are treated as negative.

Table 4: Statistics of evaluation datasets for FinNI-eval and FinCL-eval. Using the “cl100k_base” tokenizer to calculate tokens (\pm standard deviation).

Dataset	#Instance	Avg Input Tokens	Max Input Tokens	Avg Output Tokens	Max Output Tokens
FinNI-eval	6,599	986.11 \pm 835.52	13,205	138.28 \pm 302.92	5,465
FinCL-eval	52,572	61.11 \pm 44.60	750	25.00 \pm 213.77	37

3.3.2 FinNI-eval Dataset Construction

Based on the annotation procedure described in Section 3.3.1, we construct the FinNI-eval dataset, comprising 6,599 samples derived from the annotated sentences and tables. Specifically, the dataset comprises two components: the input block and the answer block. The input block includes a task instruction and the input content. The instruction defines the FinNI task by specifying the responsibilities of the LLMs, the definitions of entity types, and the rules to be followed. The answer block is formatted as a list aligned with the input block. The answer is a JSON list of all identified entity values and types or an empty JSON list. The specific prompt template is shown in the Appendix H.1.

FinNI-eval Dataset
Instruction: <FinNI Task Instruction> Input: <Sentences or Table> Answer: {"results": [{"value": <numeric entity>, "type": <entity type>}]} or {"results": []}

3.3.3 FinCL-eval Dataset Construction

Following Section 3.1, we further construct a FinCL-eval dataset that includes 52,572 query-answer pairs for numerical entity normalization. In addition, we build a US-GAAP taxonomy database containing 17,388 unique financial concepts. To create the FinCL-eval dataset, we utilize all positive instances identified in Section 3.3.1. Each query consists of a numerical entity, its corresponding entity type, and its surrounding context, while the answer is the associated US-GAAP concept.

FinCL-eval Dataset
Query: <entity> + <entity type> + <context> Answer: <US-GAAP Concept>

3.4 Evaluation

3.4.1 Evaluation Framework

We propose a unified evaluation framework for the FINTAGGING benchmark to assess LLM performance. As shown in Figure 3, given a financial report D , the framework reformulates tagging into two subtasks, FinNI and FinCL, to generate structured triplets $\{\text{Fact}, \text{Type}, \text{Tag}\}$. This design jointly evaluates an LLM’s zero-shot ability in fact extraction and concept alignment. For fair comparison with token-classification baselines, we report macro and micro Precision, Recall, and F1 [25]. FinNI is evaluated using pair-level metrics, while FinCL uses accuracy. Full metric definitions are provided in Appendix D.

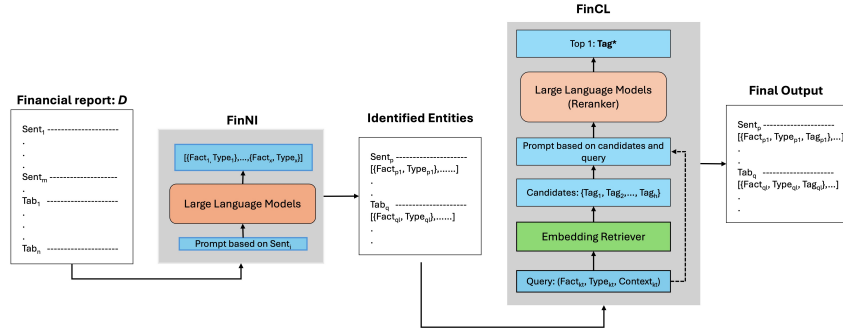


Figure 3: A unified evaluation framework on FINTAGGING benchmark.

FinNI Evaluation We use **pair-level metrics** to evaluate the extraction performance on FinNI. The objective is to evaluate the LLM’s ability to distinguish between structured and unstructured contexts and to extract financial values based on its semantic understanding and domain-specific knowledge.

FinCL Evaluation To assess whether LLMs can perform fine-grained XBRL entity normalization, we reformulate the task as a reranking problem, where LLMs are used to disambiguate and select the most appropriate taxonomy concept from a reduced candidate set, avoiding the impracticality of direct multi-thousand-way classification. To this end, we first generate embeddings for each taxonomy concept using `text-embedding-3-small`⁷, and retrieve the top- h candidate tags from the US-GAAP taxonomy \mathcal{T} based on semantic similarity. Therefore, the LLMs are required to refine the retrieved candidates by leveraging deeper contextual understanding to select the best-matched tag $\hat{c} \in \mathcal{T}_h$, formalized as:

$$f_{\text{rerank}} : (e, l, C_e, \mathcal{T}_h) \mapsto \hat{c} \quad (4)$$

3.4.2 Evaluation Models

Our goal is to evaluate the foundational capabilities of SOTA LLMs on the FINTAGGING benchmark, to better understand their strengths and limitations in financial tagging. To this end, we evaluate models across three categories using our proposed evaluation framework, as detailed in Appendix E. These include: (1) General-purpose closed-source LLM: GPT-4o model [11]; (2) General-purpose open-source LLMs including DeepSeek-V3 [13], DeepSeek-R1-Distill-Qwen-32B [6], Qwen2.5-0.5B-Instruct [34], Qwen2.5-1.5B-Instruct [34], Qwen2.5-14B-Instruct [34], Llama-3.2-3B-Instruct [9], Llama-3.1-8B-Instruct [9], and gemma-2-27b-it [26] models; and (3) Domain-specific financial LLM: Fino1-8B [19] model. In addition, we compare these LLMs with strong PLMs, including BERT-large [7], FinBERT [1], and SECBERT [14].

⁷<https://platform.openai.com/docs/models/text-embedding-3-small>

3.4.3 Evaluation setting

We use the LM Evaluation Harness [8] to build customized benchmark suites. Proprietary models such as GPT are accessed via the OpenAI API, while all other models, including DeepSeek and open-source LLMs, are evaluated through the TogetherAI API without local deployment. We standardize the input length to 2,048 tokens for FinNI and 4,096 for FinCL, with a generation limit of 1,024 tokens to support reasoning-intensive outputs. For fine-tuning strong PTMs such as BERT-large, FinBERT, and SECBERT, we curate 10 additional annual reports as training data (see Appendix F). All data follow the FiNER [14] and FNXL [23] formats. For retrieval in FinCL, we use ElasticSearch⁸ with text-embedding-3-small as the embedding model.

4 Experiment and Result

4.1 Overall Results

Table 5 presents the overall performance on the FINTAGGING benchmark. It clearly demonstrates that under our framework, LLMs can effectively handle both frequent and rare financial tags, indicating their ability to mitigate long-tail label challenges and underscoring the advantage of our information extraction and alignment formulation over traditional token-level classification approaches.

Table 5: Overall Performance. Bolded values indicate the best performance, underlined values represent the second-best, and italicized values denote the third-best performance.

Category	Models	Macro			Micro		
		P	R	F1	P	R	F1
Closed-source LLM	GPT-4o	<u>0.0764</u>	<u>0.0576</u>	<u>0.0508</u>	0.0947	0.0788	0.0860
Open-source LLMs	DeepSeek-V3	0.0813	0.0696	0.0582	0.1058	<i>0.1217</i>	<i>0.1132</i>
	DeepSeek-R1-Distill-Qwen-32B	<i>0.0482</i>	<i>0.0288</i>	<i>0.0266</i>	0.0692	0.0223	0.0337
	Qwen2.5-14B-Instruct	0.0423	0.0256	0.0235	0.0197	0.0133	0.0159
	gemma-2-27b-it	0.0430	0.0273	0.0254	0.0519	0.0453	0.0483
	Llama-3.1-8B-Instruct	0.0287	0.0152	0.0137	0.0462	0.0154	0.0231
	Llama-3.2-3B-Instruct	0.0182	0.0109	0.0083	0.0151	0.0102	0.0121
	Qwen2.5-1.5B-Instruct	0.0180	0.0079	0.0069	0.0248	0.0060	0.0096
	Qwen2.5-0.5B-Instruct	0.0014	0.0003	0.0004	0.0047	0.0001	0.0002
Financial LLM	Fino1-8B	0.0299	0.0146	0.0140	0.0355	0.0133	0.0193
Fine-tuned PLMs	BERT-large	0.0135	0.0200	0.0126	<u>0.1397</u>	0.1145	0.1259
	FinBERT	0.0088	0.0143	0.0087	<i>0.1293</i>	0.0963	0.1104
	SECBERT	0.0308	0.0483	0.0331	0.2144	0.2146	0.2145

From the macro-level perspective, which emphasizes balanced performance across both frequent and rare tags, DeepSeek-V3 and GPT-4o achieve the highest macro-F1 scores (0.0582 and 0.0508), outperforming all fine-tuned PLMs. This highlights the strong generalization of large LLMs and the effectiveness of our task design. DeepSeek-R1-Distill-Qwen-32B also achieves a solid macro-F1 (0.0266), suggesting that good architecture and pretraining can help smaller models perform well in zero-shot settings. From the micro-level perspective, which favors frequent labels, DeepSeek-V3 again performs strongly with a micro-F1 of 0.1132, ranking third overall despite no fine-tuning. GPT-4o also performs competitively with a score of 0.0860, outperforming most open-source and domain-specific models.

4.2 Subtask Results

4.2.1 FinNI subtask

Table 6 shows the performance of various LLMs on the FinNI subtask. Overall, larger models perform better at identifying numerical entities and generating structured outputs, even without financial domain training. DeepSeek-V3 achieves the highest precision and recall, outperforming all other models. In contrast, smaller models like Qwen2.5-1.5B and -0.5B score below 0.1 F1, highlighting

⁸<https://www.elastic.co/>

their limitations in zero-shot settings. Interestingly, DeepSeek-R1-Distill-Qwen-32B shows strong precision, suggesting that model design and distillation can help mitigate the impact of smaller scale.

Table 6: Performance comparison of different models on the FinNI subtask. Bolded values indicate the best performance, underlined values represent the second-best, and italicized values denote the third-best performance.

Category	Model	Precision	Recall	F1
Closed-source LLM	GPT-4o	<u>0.6105</u>	<u>0.5941</u>	<u>0.6022</u>
Open-source LLMs	Deepseek-V3	0.6329	0.8452	0.7238
	DeepSeek-R1-Distill-Qwen-32B	<i>0.5490</i>	0.2238	0.3180
	Qwen2.5-14B-Instruct	0.3632	0.0018	0.0035
	gemma-2-27b-it	0.5319	<i>0.5490</i>	<i>0.5403</i>
	Llama-3.1-8B-Instruct	0.3346	0.1746	0.2295
	Llama-3.2-3B-Instruct	0.1887	0.1794	0.1839
	Qwen2.5-1.5B-Instruct	0.1323	0.0636	0.0859
	Qwen2.5-0.5B-Instruct	0.0116	0.0027	0.0043
Financial LLM	Fino1-8B	0.3416	0.1481	0.2066

However, the domain-specific Fino1-8B model, fine-tuned on the financial reasoning QA task, does not perform competitively. This indicates that only task-specific training can effectively improve performance on FinNI; pretraining on financial corpora alone offers limited benefit if the task is not well aligned.

4.2.2 FinCL subtask

Table 7: Performance comparison of different models on the FinCL subtask. Bolded values indicate the best performance, underlined values represent the second-best, and italicized values denote the third-best performance.

Category	Model	Accuracy
Closed-source LLM	GPT-4o	<u>0.1664</u>
Open-source LLMs	Deepseek-V3	0.1715
	DeepSeek-R1-Distill-Qwen-32B	0.1013
	Qwen2.5-14B-Instruct	<i>0.1072</i>
	gemma-2-27b-it	0.1009
	Llama-3.1-8B-Instruct	0.0807
	Llama-3.2-3B-Instruct	0.0375
	Qwen2.5-1.5B-Instruct	0.0419
	Qwen2.5-0.5B-Instruct	0.0246
Financial LLM	Fino1-8B	0.0704

Table 7 shows the accuracy of different models on the FinCL subtask. Overall performance is low, highlighting the challenge of fine-grained concept linking in finance. DeepSeek-V3 achieves the highest accuracy (0.1715), followed by GPT-4o, while all other models score below 0.11. Even large open-source models like Qwen2.5-14B and Gemma-2-27B underperform, and smaller models perform near random. This reflects the difficulty of handling complex taxonomies and subtle financial semantics.

4.3 Ablation analysis

We further compare our benchmark against extreme multi-class classification settings. The prompt template is shown in Appendix H.2 and we select the best-performing model from each category (closed-source, open-source, and financial LLMs). The evaluation uses triplet-level ($\{\text{Tag}, \text{Fact}, \text{Type}\}$) Precision, Recall, and F1.

Table 8 shows that all LLMs fail completely with extreme classification, yielding zero precision, recall, and F1, confirming that a single-step choice among thousands of flat labels is not a valid LLM evaluation protocol. By decoupling extraction from concept linking and covering the full 10k+ taxonomy, FinTagging produces meaningful scores and thus offers a far more realistic test bed for future model improvements.

Table 8: Performance comparison between w/wo our evaluation framework on the FINTAGGING benchmark dataset.

Evaluation Mode	Model	Precision	Recall	F1
FinTagging	GPT-4o	0.1075	0.0972	0.1021
	Deepseek-V3	0.1229	0.1519	0.1359
	Fino1-8B	0.0283	0.0138	0.0185
Extreme Classification	GPT-4o	0	0	0
	Deepseek-V3	0	0	0
	Fino1-8B	0	0	0

5 Conclusion

This paper presents FINTAGGING, a benchmark for evaluating large language models on XBRL tagging of real-world financial reports. The task is divided into two subtasks, financial numeric identification (FinNI) and concept linking (FinCL), to enable fine-grained evaluation of both information extraction and taxonomy alignment. Results show that while LLMs generalize well to long-tail entities and perform competitively in zero-shot settings, they struggle with accurate alignment to GAAP concepts. This reveals limitations in schema-aware reasoning and highlights the need for better semantic understanding. FINTAGGING offers a foundation for advancing research in XBRL tagging and regulatory reporting. The limitations of our work and directions for future research are discussed in Appendix A.

References

- [1] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.
- [2] Matthew Bovee, Alexander Kogan, Kay Nelson, Rajendra P Srivastava, and Miklos A Vasarhelyi. Financial reporting and auditing agent with net knowledge (fraank) and extensible business reporting language (xbri). *Journal of Information Systems*, 19(1):19–41, 2005.
- [3] Zhiyuan Cao, Vipina K Keloth, Qianqian Xie, Lingfei Qian, Yuntian Liu, Yan Wang, Rui Shi, Weipeng Zhou, Gui Yang, Jeffrey Zhang, et al. The development landscape of large language models for biomedical applications. *Annual Review of Biomedical Data Science*, 8, 2025.
- [4] Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*, 2021.
- [5] Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. Convinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *arXiv preprint arXiv:2210.03849*, 2022.
- [6] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
- [8] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024.
- [9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [10] Shijie Han, Haoqiang Kang, Bo Jin, Xiao-Yang Liu, and Steve Yang. Xbrl-agent: Leveraging large language models for financial report analysis. In *Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF ’24)*, 2024.

- [11] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [12] Rik Koncel-Kedziorski, Michael Krumbick, Viet Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. Bizbench: A quantitative reasoning benchmark for business and finance. *arXiv preprint arXiv:2311.06602*, 2023.
- [13] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [14] Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. Finer: Financial numeric entity recognition for xbrl tagging. *arXiv preprint arXiv:2203.06482*, 2022.
- [15] Chenxi Ma, Zhen Huang, Jiaxin Wei, and Xu Sun. Label semantics enhanced financial entity recognition. *arXiv preprint arXiv:2203.06482*, 2022.
- [16] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.
- [17] Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, et al. Ectsum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. *arXiv preprint arXiv:2210.12467*, 2022.
- [18] Xueqing Peng, Triantafillos Papadopoulos, Efstathia Soufleri, Polydoros Giannouris, Ruoyu Xiang, Yan Wang, Lingfei Qian, Jimin Huang, Qianqian Xie, and Sophia Ananiadou. Plutus: Benchmarking large language models in low-resource greek finance, 2025.
- [19] Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Jimin Huang, and Qianqian Xie. Fino1: On the transferability of reasoning enhanced llms to finance. *arXiv preprint arXiv:2502.08127*, 2025.
- [20] Jim Richards, Barry Smith, and Ali Saeedi. An introduction to xbrl. *Available at SSRN 1007570*, 2006.
- [21] Rachit Saini, Ankit Gupta, and Harshil Singh. Galaxc: Graph neural networks for extreme classification in financial text. *arXiv preprint arXiv:2104.05709*, 2021.
- [22] Agam Shah, Ruchit Vithani, Abhinav Gullapalli, and Sudheer Chava. Finer: Financial named entity recognition dataset and weak-supervision model. *arXiv preprint arXiv:2302.11157*, 2023.
- [23] Soumya Sharma, Subhendu Khatuya, Manjunath Hegde, Afreen Shaikh, Koustuv Dasgupta, Pawan Goyal, and Niloy Ganguly. Financial numeric extreme labelling: A dataset and benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3550–3561, 2023.
- [24] Soumya Sharma, Tapas Nayak, Arusarka Bose, Ajay Kumar Meena, Koustuv Dasgupta, Niloy Ganguly, and Pawan Goyal. Finred: A dataset for relation extraction in financial domain. In *Companion Proceedings of the Web Conference 2022*, pages 595–597, 2022.
- [25] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.
- [26] Gemma Team. Gemma. 2024.
- [27] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*, 2019.
- [28] Richard Zhe Wang. Standardizing xbrl financial reporting tags with natural language processing. *Available at SSRN 4613085*, 2023.
- [29] Yan Wang, Lingfei Qian, Xueqing Peng, Jimin Huang, and Dongji Feng. Ordrankben: A novel ranking benchmark for ordinal relevance in nlp, 2025.
- [30] Yan Wang, Jian Wang, Huiyi Lu, Bing Xu, Yijia Zhang, Santosh Kumar Banbhrani, Hongfei Lin, et al. Conditional probability joint extraction of nested biomedical events: design of a unified extraction framework based on neural networks. *JMIR Medical Informatics*, 10(6):e37804, 2022.

- [31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [32] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*, 2019.
- [33] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*, 2023.
- [34] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.
- [35] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*, 2021.

A Limitations

In this work, we propose FINTAGGING and conduct a comprehensive analysis of LLM performance on the task. However, several limitations remain. First, we have not yet evaluated a broader range of models, including recent releases such as GPT-4.1, LLaMA-4, and Qwen-3. We are currently conducting these experiments and will report the results in future work. Second, although we provide detailed empirical results, we did not perform statistical significance testing, as evaluating the complete benchmark, especially for models like GPT and DeepSeek, requires substantial time. This analysis is also in progress and will be included in future updates. Third, we collected only a small dataset for fine-tuning PLMs and did not construct task-specific training data for LLMs. Future work may explore building LLM-based financial tagging agents through targeted fine-tuning to further enhance performance.

B Literature Review

The XBRL provides a comprehensive taxonomy for financial reporting, encompassing thousands of detailed tags corresponding to concepts within financial statements. Applying NER to assign XBRL tags is an emerging yet challenging area.

B.1 XBRL Tagging benchmark

FiNER systematically benchmarked several neural architectures on the finer-139 dataset to address numeric-heavy XBRL tagging [14]. The initial experiments showed that standard BERT underperforms due to subword fragmentation, then the authors introduced pseudo-token strategies replacing numerals with [NUM] or [SHAPE] tokens to stabilize label assignment across fragmented numeric spans. These strategies, combined with domain-specific pretraining on SEC-BERT, significantly improved tagging performance, reaching 82.1 micro-F1 without the need for computationally expensive CRF layers. Their experiments demonstrated that subword-aware models with numeric-aware pseudo-tokens outperform both word-level BiLSTMs and vanilla BERT, particularly in numeric-heavy contexts, and avoid nonsensical label sequences. FNXL extended this benchmarking paradigm to a much larger label space of 2,794 US-GAAP tags, reframing the task as an extreme classification problem [23]. They compared the FiNER sequence labeling approach with a two-step pipeline that first identifies numeric spans and then assigns labels using AttentionXML. While FiNER achieved stronger micro-F1 (75.84), reflecting better performance on frequent tags, AttentionXML outperformed FiNER in macro-F1 (47.54), highlighting its strength in predicting infrequent, tail-end labels. FNXL further evaluated both models under a Hits@k setting, confirming that label recommendations from the AttentionXML pipeline could substantially reduce manual effort and maintain high inter-annotator agreement. Together, these benchmarks reveal the need for context-aware reasoning and label-ranking mechanisms in realistic XBRL tagging scenarios.

B.2 XBRL Tagging Methods

The previous studies also explored the approaches address scalability, semantic ambiguity, and reasoning gaps in XBRL tagging to improve the performance. Saini et al. [21] proposed GalaXC is a graph-based extreme classification framework that jointly learns over document-label graphs with per-label attention across multi-hop neighborhoods. By integrating label metadata and transitive label correlations, GalaXC outperformed leading deep classifiers by up to 18% in micro-F1 on standard benchmarks and achieved 25% gains in warm-start scenarios where partial labels are available. Moreover, Wang et al. [28] addressed the practical challenge of custom tag standardization through a semantic similarity pipeline that leverages TF-IDF, Word2Vec, and FinBERT embeddings. Although unsupervised, the method was tested across nearly 200,000 custom tags from SEC filings between 2009 and 2022, and showed strong alignment performance, with vector-based mappings identifying viable standard tag candidates for a substantial proportion of non-compliant elements—offering a low-cost, interpretable solution for downstream financial analysis. Shifting focus from classification to comprehension, XBRL-Agent evaluated the capabilities of large language models to reason over full XBRL reports [10]. The authors introduced two task types—domain taxonomy understanding and numeric reasoning and found that base LLMs often hallucinated or misinterpreted financial content. To overcome these issues, XBRL-Agent incorporated retrieval-augmented generation (RAG) and symbolic calculators within an LLM-agent framework. The enhanced system achieved a 17% accuracy gain on domain query tasks and a 42% boost on numeric reasoning queries compared to base LLMs, validating the utility of modular tool augmentation. These improvements enabled reliable multi-step reasoning over complex disclosures such as debt instruments and derivative gains, which are difficult to capture using span-level classifiers. Collectively, these works broaden the methodological landscape of XBRL tagging from graph-based label propagation and embedding-based normalization to LLM-driven report analysis and point to a hybrid future where structured priors and reasoning tools jointly support accurate, scalable financial information extraction.

B.3 Financial Evaluation Benchmarks

In parallel to XBRL-specific advances, the financial NLP community has developed comprehensive benchmarks to assess broader capabilities in information extraction, numerical reasoning, and document understanding. FINER-ORD [22] introduced a high-quality, domain-specific NER dataset annotated over financial news, emphasizing general entity types like persons, organizations, and locations. While not numerically focused like FINER-139, it highlights the lexical diversity of financial discourse and establishes a strong baseline for testing pretrained and zero-shot LLMs in real-world financial NER scenarios. FinQA [4] pushed toward explainable QA by pairing expert-written questions with annotated multi-step reasoning programs derived from earnings reports. ConvFinQA [5] extended this challenge to conversational contexts, simulating real-world question flows over sequential financial queries. TAT-QA [35] focused on hybrid tabular-text reasoning and required models to align cell values and document narratives, often involving aggregation, comparison, and unit-scale interpretation. Pixiu [33] introduced a broader evaluation framework by releasing FinMA, a financial LLM instruction-tuned across five tasks, and assessing it on a new benchmark covering sentiment classification, QA, summarization, NER, and stock prediction. BizBench [12] framed financial QA as program synthesis over realistic, multi-modal contexts, integrating reasoning, code generation, and domain knowledge into a single evaluation pyramid. While these benchmarks highlight the growing ability of models to integrate structured and unstructured financial data, they overlook taxonomy-driven fact alignment and do not support the structured output formats required for XBRL tagging.

C The Statistics of the Company

Table 9 lists the 30 publicly traded companies whose 2024 annual 10-K filings were used in our study. For each company, we report its Central Index Key (CIK), legal name, stock ticker symbol, and the fiscal year-end filing date. These reports were retrieved from the SEC EDGAR system and conform to the XBRL standard. The selected filings span a diverse set of industries and ensure a representative sample for analyzing structured and unstructured financial disclosures. This company list complements the dataset statistics summarized in Table 2.

Table 9: Company filing summary (2024) for benchmark data

CIK	Company Name	Ticker	Filing Date
0001163739	NABORS INDUSTRIES LTD	NBR	2024-12-31
0001418819	Iridium Communications Inc.	IRDM	2024-12-31
0001993004	NorthWestern Energy Group, Inc.	NWE	2024-12-31
0000024741	CORNING INC /NY	GLW	2024-12-31
0000079282	BROWN & BROWN, INC	BRO	2024-12-31
0001364479	HERC HOLDINGS INC	HRI	2024-12-31
0000936340	DTE ENERGY CO	DTM	2024-12-31
0000085535	ROYAL GOLD INC	RGLD	2024-12-31
0000086312	TRAVELERS COMPANIES, INC.	TRV	2024-12-31
0001968915	PHINIA INC.	PHIN	2024-12-31
0000093751	STATE STREET CORP	STT	2024-12-31
0000106640	WHIRLPOOL CORP	WHR	2024-12-31
0000888491	OMEGA HEALTHCARE INVESTORS INC	OHI	2024-12-31
0001616862	Axalta Coating Systems Ltd.	AXTA	2024-12-31
0000922224	PPL Corp	PPL	2024-12-31
0000066756	ALLETE INC	ALE	2024-12-31
0001411207	Allison Transmission Holdings Inc	ALSN	2024-12-31
0001478242	IQVIA HOLDINGS INC.	IQV	2024-12-31
0001544400	NFinTi inc.	NFTN	2024-12-31
0000851205	COGNEX CORP	CGNX	2024-12-31
0001318220	Waste Connections, Inc.	WCN	2024-12-31
0000048898	HUBBELL INC	HUBB	2024-12-31
0000937098	TRINET GROUP, INC.	TNET	2024-12-31
0000910108	LXP Industrial Trust	LXP	2024-12-31
0001932393	GE HealthCare Technologies Inc.	GEHC	2024-12-31
0001770787	10x Genomics, Inc.	TXG	2024-12-31
0001026214	FEDERAL HOME LOAN MORTGAGE CORP	FMCC	2024-12-31
0000004904	AMERICAN ELECTRIC POWER CO INC	AEP	2024-12-31
0001637459	Kraft Heinz Co	KHC	2024-12-31
0001713445	Reddit, Inc.	RDDT	2024-12-31

D Evaluation Metrics

To provide a fair evaluation of overall benchmark performance, we adopt a set of metrics, focusing primarily on macro-level and micro-level evaluation strategies inspired by the previous work [23]. **Macro-level** evaluation computes precision, recall, and F1 scores independently for each BIO-concept label derived from the US-GAAP taxonomy, and then averages them without weighting. This ensures that each concept, including rare or infrequent ones, contributes equally to the final score—making it especially suitable for domains with skewed label distributions. In contrast, **micro-level** evaluation aggregates token-level true positives, false positives, and false negatives across all labels before computing precision, recall, and F1. This approach emphasizes the model’s overall tagging accuracy by treating every token equally, and thus better reflects performance on frequent concepts. Together, these two metrics provide a balanced view of both per-concept performance and overall tagging quality.

For the FinNI subtask, the objective is to extract correct **(entity, type)** pairs, that is **(Fact, Type)**, from the financial document. Let $\mathcal{G} = (e_i, l_i)$ denote the set of ground-truth (entity, type) pairs, and $\mathcal{P} = (e'_i, l'_i)$ denote the set of predicted (entity, type) pairs. We evaluate the performance based on the following metrics:

$$\text{Precision} = \frac{|\mathcal{G} \cap \mathcal{P}|}{|\mathcal{P}|} \quad (5)$$

$$\text{Recall} = \frac{|\mathcal{G} \cap \mathcal{P}|}{|\mathcal{G}|} \quad (6)$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

where $(e_i, l_i) = (e'_j, l'_j)$ if and only if both the entity span e and its assigned type l exactly match.

For the FinCL subtask, Given a set of queries $\mathcal{Q} = \{q_1, q_2, \dots, q_N\}$, where each q_i is associated with a ground-truth concept c_i^* and a predicted concept \hat{c}_i , the accuracy is defined as:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \delta(\hat{c}_i, c_i^*) \quad (8)$$

where $\delta(\hat{c}_i, c_i^*) = 1$ if $\hat{c}_i = c_i^*$, and 0 otherwise.

E Evaluation Models Details

Table 10 provides an overview of the models evaluated in this study, categorized by openness, domain specialization, and architectural foundation. The evaluation covers a diverse range of models:

- **Closed-source LLMs:** We include GPT-4o [11], accessed via OpenAI’s API, as a representative of cutting-edge proprietary models with demonstrated performance across a variety of NLP tasks. Although model size details are undisclosed, GPT-4o serves as an upper-bound reference in our benchmark.
- **Open-source LLMs:** This group encompasses recent, high-performing open models such as DeepSeek-V3 (685B) [13], DeepSeek-R1-Distill-Qwen (32B) [6], and multiple variants of Qwen2.5-Instruct [34](ranging from 0.5B to 14B). We also include LLaMA-3.2 and 3.1 variants [9] (3B and 8B), as well as Google’s Gemma-2-27B [26], to ensure architectural diversity and scalability comparison. These models are primarily instruction-tuned and optimized for general-purpose NLP tasks.
- **Financial-specific LLMs:** We evaluate Fino1-8B [19], a domain-specialized model trained on financial corpora, designed to better capture the terminology and structure unique to financial disclosures. This category allows us to assess the benefits of domain adaptation in complex tagging and reasoning tasks.
- **Pretrained Language Models (PLMs):** To establish strong baselines, we include non-generative encoder models: BERT-large [7], FinBERT [1], and SECBERT [14]. These models have been widely used in prior financial NLP tasks and allow for a comparative analysis between generative LLMs and traditional pretrained models in terms of domain understanding and structured output capability.

Together, these models offer a comprehensive evaluation spectrum, from general-purpose to domain-specific, encoder-based to decoder-based, and open to closed source, facilitating an in-depth assessment of their performance across our proposed benchmark tasks.

Table 10: Model Categories and Corresponding Repositories

Model	Size	Source
Closed-source Large Language Models		
GPT-4o	-	gpt-4o-2024-08-06
Open-source Large Language Models		
DeepSeek-V3	685B	deepseek-ai/DeepSeek-V3
DeepSeek-R1-Distill-Qwen	32B	deepseek-ai/DeepSeek-R1-Distill-Qwen-32B
Qwen2.5-Instruct	0.5B, 1.5B, 14B	Qwen/Qwen2.5-*B-Instruct
Llama-3.2-Instruct	3B	meta-llama/Llama-3.2-3B-Instruct
Llama-3.1-Instruct	8B	meta-llama/Llama-3.1-8B-Instruct
gemma-2-27b-it	27B	google/gemma-2-27b-it
Financial-specific Large Language Models		
Fino1	8B	TheFinAI/Fino1-8B
Pretrained Language Models		
BERT-large	~340M	google-bert/bert-large-uncased
FinBERT	~110M	ProsusAI/finbert
SECBERT	~110M	nlpaueb/sec-bert-base

F The details for the fine-tuning PTMs

F.1 Training data collection and processing

Similar to the collection process for the FINTAGGING benchmark data, we gathered an additional 10 annual 10-K financial reports filed with the SEC for the period from February 13, 2024, to February 13, 2025, as summarized in Table 11. These reports contain a total of 33,848 standard taxonomy-tagged facts. Using BeautifulSoup to parse these documents, we identified 22,847 narrative sentences (approximately 5.5 million characters) and 1,236 financial tables. The companies included in this dataset follow the XBRL standard, ensuring comprehensive coverage for training PTMs.

Table 11: Financial report statistics summary for raw training data

Item	Information
Report type	10-K
Period	2024-02-13 to 2025-02-13
#Company	10
#Sentence	22,847
#Table	1,236
#Characters	5,539,198
#Standard Tags	33,848

After collection, we employed the same procedure to filter texts and tables, subsequently annotating numerical entities, entity types, and US-GAAP tags (concepts). Finally, as detailed in Table 12, we generated a total of 1,116 sentences and 953 tables as the training set for PTMs. Specifically, the sentence-level data consists of 558 positive and 558 negative instances, averaging approximately 84.24 tokens (± 69.29), with 1.22 annotated entities and concepts per sentence. The table-level data comprises 594 positive and 359 negative instances, with a significantly higher average of 1,281.86 tokens ($\pm 6,438.37$), and approximately 25 entities and concepts annotated per table. Overall, the annotated dataset includes 25,199 entities, covering 1,435 unique US-GAAP concepts.

Table 12: The statistical information for training data. Using the “cl100k_base” tokenizer to calculate tokens (\pm standard deviation).

Structure	Pos/Neg	#Instance	Avg. Tokens/S	Avg. Entities/S	Avg. Concepts/S	Total Entities	Unique Concepts
Sentence	Positive	558	84.24 ± 69.29	1.22 ± 1.78	1.22 ± 1.78	25,199	1,435
	Negative	558					
Table	Positive	594	1281.86 ± 6438.37	25.00 ± 213.77	25.00 ± 213.77		
	Negative	359					

However, to align with the extreme classification format used in previous XBRL tagging benchmarks, we directly adopt the US-GAAP tags as entity labels, annotating each token in sentences and tables using the BIO scheme. Specifically, B denotes the beginning of an entity phrase, I marks the continuation (inside) of an entity phrase,

and 0 indicates tokens outside of any entity. As shown in Figure 4, "4.9" and "4.5" are single-token numerical entities labeled only with a B prefix (e.g., "B-us-gaap:AccountsReceivableNetNoncurrent"). To comprehensively cover all US-GAAP tags, we combine the entire set of 17,388 tags from the US-GAAP 2024 taxonomy with the BIO labeling scheme to construct an extreme classification label space, resulting in 34,777 unique entity labels ($2 \times 17388 + 1$).

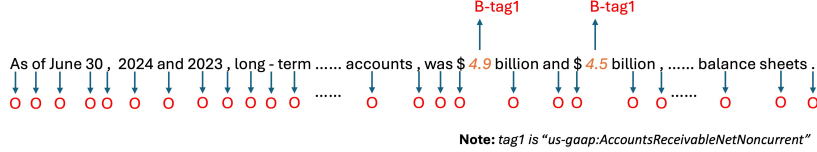


Figure 4: An example for training set annotation with BIO scheme.

After constructing the training set, we reconstruct the testing set from the original benchmark dataset. The training settings are detailed below.

F.2 Training settings

We fine-tune three pretrained models, BERT-large [7], FinBERT [1], and SECBERT [14], on our training set using the HuggingFace Transformers library. All models are trained with a batch size of 4, a learning rate of $3e-5$, and for 20 epochs. Optimization is performed using AdamW without gradient accumulation or early stopping. Token classification heads are randomly initialized and trained jointly with the base encoder. Input sequences are tokenized with a maximum length of 512, and labels are aligned at the sub-token level following the BIO tagging scheme. Loss is computed only on the first sub-token of each word to avoid misalignment bias.

Training is conducted on two NVIDIA A5000 GPUs (24GB each) using data parallelism for 24 hours. All other hyperparameters follow the default settings in the HuggingFace Trainer API. Models are evaluated using the checkpoint from the final epoch. All experiments are run under a fixed random seed to ensure reproducibility.

G Retrieval Results in FinCL subtask

We investigate the impact of different context construction strategies for the queried entity at the **retrieval stage**. We consider two approaches: Fixed-Window Context (FWC) and Structure-Aware Context (SAC). In the FWC strategy, context is constructed by extracting a fixed window of 50 characters before and after the entity mention, regardless of whether the entity appears in a sentence or a table. In contrast, the SAC strategy builds context based on the structural location of the entity: if the entity appears in a sentence, the entire sentence is used; if the entity appears in a table, we linearize the entire row into a Markdown-style key-value format to serve as the context.

Table 13: The Acc@k performance of retrieval for different context construction strategies on the FinCL task. Here, FWC denotes the fixed window-based strategy, while SAC denotes the structure-aware approach that respects sentence or row boundaries.

Strategy	Structure	Acc@1	Acc@10	Acc@20	Acc@50	Acc@100	Acc@150	Acc@200
FWC	Sentence	0.0677	0.2594	0.3554	0.4650	0.5515	0.6002	0.6404
	Table	0.000	0.0030	0.0031	0.0036	0.0050	0.0064	0.0097
	Overall	0.0055	0.0237	0.0316	0.0409	0.0492	0.0544	0.0608
SAC	Sentence	0.0689	0.2643	0.3521	0.4711	0.5611	0.6096	0.6522
	Table	0.0147	0.0867	0.1240	0.1855	0.2353	0.2632	0.2866
	Overall	0.0191	0.1011	0.1424	0.2086	0.2617	0.2913	0.3163

From Table 13, we found the SAC strategy consistently outperforms FWC, particularly in the table context, highlighting the importance of aligning the context window with the underlying structural unit. For **sentence-based entities**, both strategies yield comparable results, with SAC achieving slightly higher Acc@1 (0.0689 vs. 0.0677) and showing marginal improvements across all cutoff values. This suggests that even for relatively unstructured text, preserving sentence boundaries may provide minor benefits over a fixed-length window.

In contrast, for **table-based entities**, the performance gap is substantial. SAC achieves significantly higher retrieval accuracy (e.g., Acc@100 of 0.2353 vs. 0.0050), indicating that row-level context is far more informative than arbitrary character windows when dealing with tabular structures. FWC performs poorly in this setting, likely due to the fragmented and semantically sparse nature of partial table text.

When aggregating results across both structures, SAC outperforms FWC by a wide margin at all retrieval depths (e.g., Acc@200 of 0.3163 vs. 0.0608). These results underscore the importance of structure-aware context construction, especially in scenarios where inputs span multiple formats such as sentences and tables.

H Prompt Templates

H.1 Template for FinNI subtask

Prompt Template

You are a financial information extraction expert specializing in identifying financial numerical entities in XBRL reports. Your task is to extract all such numerical entities from the provided text or serialized `<table></table>` data and classify them into one of five categories:

- "integerItemType": Counts of discrete items, such as the number of employees or total transactions.
- "monetaryItemType": Financial amounts expressed in currency, such as revenue, profit, or total assets.
- "perShareItemType": Per-share values, such as earnings per share (EPS) or book value per share.
- "sharesItemType": Counts of shares, such as outstanding shares or ownership stakes.
- "percentItemType": Ratios or percentages, such as tax rates, growth rates, or discount rates, usually expressed with a percentage symbol ("%").

Important Instructions:

- (1) Financial numerical entities are not limited to Arabic numerals (e.g., 10,000). They may also appear in word form (e.g., "ten million"), which must be correctly identified and converted into standard numerical format.
- (2) Not all numbers in the text should be extracted. Only those that belong to one of the five financial entity categories above should be included. Irrelevant numbers (such as phone numbers, dates, or general IDs) must be ignored.
- (3) If a number is followed by a magnitude term (e.g., Hundred, Thousand, Million, Billion), do not expand it into the full numerical value.
 - * "Two hundred" -> Extract only "two", not "200".
 - * "10.6 million" -> Extract only "10.6", not "10,600,000".
- (4) Standardize numerical formatting by removing currency symbols (e.g., "USD"), percentage signs ("%"), and commas (",") while preserving the numeric value. These elements must be removed to ensure consistency.
- (5) Output the extracted financial entities in JSON list format without explanations, structured as follows: {"result":[{"Fact": <Extracted Numerical Entity>, "Type": <Identified Entity Type>}]}

Input: {text/table}

Output:

H.2 Template for ablation

Prompt Template

You are an XBRL tagging expert specializing in annotating financial numerical facts in XBRL reports.
Your task is to (1) extract all such numerical entities from the provided text or serialized <table></table> data, (2) classify them into one of five categories, and (3) assign an appropriated US-GAAP tag to each entity.

Categories:

- "integerItemType": Counts of discrete items, such as the number of employees or total transactions.
- "monetaryItemType": Financial amounts expressed in currency, such as revenue, profit, or total assets.
- "perShareItemType": Per-share values, such as earnings per share (EPS) or book value per share.
- "sharesItemType": Counts of shares, such as outstanding shares or ownership stakes.
- "percentItemType": Ratios or percentages, such as tax rates, growth rates, or discount rates, usually expressed with a percentage symbol ("%").

US-GAAP tags:

- A US-GAAP tag is a standardized semantic label used in XBRL filings to identify specific financial concepts defined by the U.S. Generally Accepted Accounting Principles (GAAP). Each tag represents a distinct accounting item and enables consistent, machine-readable financial reporting.
- Examples: "us-gaap:AssetsCurrentAbstract", "us-gaap:AccruedInsuranceNoncurrent".

Important Instructions:

- (1) Financial numerical entities are not limited to Arabic numerals (e.g., 10,000). They may also appear in word form (e.g., "ten million"), which must be correctly identified and converted into standard numerical format.
- (2) Not all numbers in the text should be extracted. Only those that belong to one of the five financial entity categories above should be included. Irrelevant numbers (such as phone numbers, dates, or general IDs) must be ignored.
- (3) If a number is followed by a magnitude term (e.g., Hundred, Thousand, Million, Billion), do not expand it into the full numerical value.
 - * "Two hundred" -> Extract only "two", not "200".
 - * "10.6 million" -> Extract only "10.6", not "10,600,000".
- (4) Standardize numerical formatting by removing currency symbols (e.g., "USD"), percentage signs ("%"), and commas (",") while preserving the numeric value. These elements must be removed to ensure consistency.
- (5) You should assign the most appropriate US-GAAP tag to each identified entity based on your internal understanding of the 2024 US-GAAP taxonomy.
- (6) Output the extracted financial entities in JSON list format without explanations, structured as follows: {"result":[{"Fact": <Extracted Numerical Entity>, "Type": <Identified Entity Type>, "Tag": <Assigned US-GAAP tag>}]}

Input: {text/table}

Output:

I Potential Societal Impacts of the Work

I.1 Potential Positive Societal Impacts

This work introduces FINTAGGING, a full-scope, structure-aware benchmark for evaluating large language models on XBRL-based financial reporting. By supporting realistic tagging scenarios grounded in US-GAAP taxonomy, this benchmark:

- Promotes automation and transparency in financial disclosure, which can reduce manual errors and increase the reliability of corporate filings.
- Enables more accurate, scalable tools for financial regulators, auditors, and analysts, potentially improving oversight and decision-making in global markets.
- Encourages the development of LLMs with stronger schema-aware reasoning, a critical step toward trustworthy AI systems in high-stakes financial environments.

Through open access to annotated datasets and evaluation code, FINTAGGING also advances reproducible research and supports the broader community in benchmarking financial language models responsibly.

I.2 Potential Negative Societal Impacts

While FINTAGGING provides valuable tools for model evaluation, several risks should be noted:

- **Misuse and Misinterpretation:** Benchmark performance may be misinterpreted as real-world readiness. Deployment of LLMs without sufficient regulatory safeguards could lead to incorrect tagging, misreporting, or financial misinformation.
- **Coverage Bias:** The benchmark focuses on US-GAAP financial reports and may underrepresent other accounting standards or financial contexts, potentially limiting its applicability in non-U.S. jurisdictions.
- **Overreliance on Automation:** Improved tagging accuracy may incentivize overautomation in financial reporting pipelines, increasing the risk of undetected errors if human oversight is reduced.

To mitigate these concerns, we recommend that future work includes broader financial contexts and integrates FinTagging into a framework with human-in-the-loop validation, fairness checks, and error auditing. Ongoing collaboration with domain experts and regulators is also essential.