國立臺灣大學管理學院資訊管理學系

碩士論文

Department of Information Management
College of Management
National Taiwan University
Master's Thesis

# 金融數值實體理解：任務、資料集與方法

# Financial Numerical Entity Understanding: Novel Tasks, Datasets, and Approaches

李宜蓁

Yi-Jhen Li

指導教授：盧信銘 博士

Advisor: Hsin-Min Lu, Ph.D.

中華民國 114 年 7 月

July, 2025

doi:10.6342/NTU202501600

# 誌謝

完成這篇碩士論文的過程中，收到許多人的指導、鼓勵與陪伴，才能順利走到最後，感謝所有在這段旅程中給予我支持與幫助的人。首先，感謝我的指導教授盧信銘老師，謝謝老師在研究過程中耐心指導，提供專業的建議。另外也感謝顏如君老師及謝昇峯老師，撥冗擔任口試委員並提出寶貴的回饋，幫助我更加完善研究內容。此外，要特別謝謝一路以來陪伴我的家人，謝謝你們的理解與支持，讓我能無後顧之憂地專注於學業與研究。也謝謝在這段日子裡陪我走過高低起伏的朋友與實驗室夥伴們，因為有你們在生活與研究中帶來的溫暖與歡笑，讓碩士生活增添了更多色彩。最後，也謝謝自己在這段研究旅程中堅持下來，完成了曾經不敢想像能做到的目標。再次向所有曾經幫助與支持我的人致上最誠摯的謝意，期許未來能夠帶著這份感激，繼續努力前行。



doi:10.6342/NTU202501600

# 中文摘要

隨著 inline XBRL (iXBRL) 在財務報告中被廣泛與強制使用，能夠提升效率、一致性及法規遵循的自動化標記需求日益增加。人工標記不僅耗時費力，且容易出錯。然而，現有資料集在屬性涵蓋範圍及文本資訊上仍有不足，限制了模型效能及其在真實應用場景中的適用性。為解決上述問題，本研究從美國證券交易委員會（SEC）申報文件中建構出一套大規模 iXBRL 標記資料集，包含 660 萬筆句子。每筆資料包含多個屬性，如標籤名稱（tag name）、財務數值（fact value）、時間（time）等，並附有財務報表資訊及句子上下文資訊。本研究亦將標記任務重新定義為多屬性預測問題，以更真實地反映實際財務標記的複雜性。我們提出三階段任務設計，提升資料集在不同模型與應用場景中的靈活性與可用性。本研究比較兩種方法：基於 BERT 架構的多任務學習（multi-task learning, MTL）模型，以及大型語言模型（LLMs）進行的 few-shot 提示學習（prompting）。MTL 模型在多個屬性上展現優異的預測能力，於標籤、時間、數量級與正負屬性的加權 F1 分數分別達到 0.82、0.90 及 0.99。相較之下，few-shot LLMs 表現則相對較差。這一結果凸顯了大型語言模型和 few-shot 在財務領域的限制。此外，本研究所建資料集亦能協助發現過往申報文件中的標記不一致或潛在錯誤，展現其作為訓練資源與稽核工具的價值。從管理實務觀點而言，本研究提出的資料集與自動標記框架可大幅降低人工工作量，提升財務揭露的可靠性，並促進後續應用，例如對股東會資料、財報摘要，或在 XBRL 強制採用前發布之歷史財報的結構化分析。

關鍵詞：XBRL、iXBRL、財務資料集、屬性預測、多任務學習



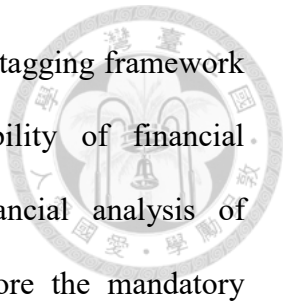
doi:10.6342/NTU202501600

# ABSTRACT

The growing adoption of inline XBRL (iXBRL) in financial reporting has increased the need for automation to improve efficiency, consistency, and compliance. Manual tagging is often labor-intensive and error-prone, especially in corporate settings where financial filings must meet strict regulatory standards. While automated iXBRL tagging has shown promise, existing datasets lack comprehensive attribute coverage and contextual information, limiting model performance and real-world applicability. To address these limitations, this study constructs a large-scale iXBRL tagging dataset containing 6.6 million sentences from SEC filings. Each instance is annotated with multiple attributes, such as tag name, fact value, and time, as well as document metadata and surrounding sentence context. We reformulate the tagging task as a multi-attribute prediction problem, which better reflects the complexity of real-world financial reporting. A three-stage task design is proposed to improve the flexibility and usability of the dataset for various modeling approaches and applications. To establish performance baselines, we evaluate two methods: a multi-task learning (MTL) model based on a BERT architecture and few-shot prompting using large language models (LLMs). The MTL model demonstrates strong predictive capabilities across attributes, achieving weighted F1 scores of 0.82 for tag, 0.90 for time, and 0.99 for both scale and sign. By contrast, few-shot LLMs perform relatively worse. These findings reveal the current limitations of prompt-based approaches and highlight opportunities for future improvement through domain-adaptive pretraining and advanced prompting strategies. Additionally, the dataset helped uncover annotation inconsistencies and potential errors in previous filings, highlighting its value as both a training resource and an auditing aid.



doi:10.6342/NTU202501600



From a managerial perspective, the proposed dataset and automated tagging framework can significantly reduce human workload, enhance the reliability of financial disclosures, and enable downstream applications, such as financial analysis of shareholder meeting materials or financial reports published before the mandatory XBRL adoption.

Keywords: XBRL, iXBRL, Financial Dataset, Attribute Prediction, Multi-task Learning



doi:10.6342/NTU202501600

# TABLE OF CONTENTS

誌謝 ..... i

中文摘要 ..... ii

ABSTRACT ..... iii

TABLE OF CONTENTS ..... v

LIST OF FIGURES ..... ix

LIST OF TABLES ..... x

Chapter 1 Introduction ..... 1

Chapter 2 Literature Review ..... 5

2.1 XBRL ..... 5

2.1.1 Overview of XBRL ..... 5

2.1.2 XBRL Structure and Component ..... 6

2.1.2.1 Reported Facts ..... 6

2.1.2.2 XBRL Instance Document ..... 7

2.1.2.3 Concepts ..... 8

2.1.2.4 XBRL Taxonomy ..... 9

2.1.3 XBRL Adoption ..... 10

2.1.3.1 Global Adoption of XBRL ..... 10

2.1.3.2 Advantages of XBRL Adoption ..... 11

2.1.3.3 Challenge of XBRL Adoption ..... 12

2.2 Inline XBRL ..... 12

2.2.1 Overview of iXBRL ..... 13

2.2.2 iXBRL Structure and Tagging ..... 13



doi:10.6342/NTU202501600

2.2.2.1 iXBRL Report Structure ..... 13
2.2.2.2 iXBRL Tagging ..... 15
2.2.2.2.1 Format ..... 16
2.2.2.2.2 Scale ..... 16
2.2.2.2.3 Sign ..... 16
2.2.2.2.4 Textual data ..... 17
2.2.3 iXBRL Adoption ..... 18
2.2.3.1 Global Adoption of iXBRL ..... 18
2.2.3.2 Advantages of iXBRL Adoption ..... 19
2.3 NLP, LLM, and XBRL ..... 20
2.3.1 LLM-based Analysis of XBRL ..... 20
2.3.2 Automated Tagging in XBRL ..... 21
2.4 Multi-task Learning ..... 22
2.5 Research Gap ..... 23
Chapter 3 Methodology ..... 27
3.1 Dataset ..... 27
3.1.1 Dataset Overview ..... 27
3.1.2 Dataset Construction Process ..... 28
3.1.3 Data Structure ..... 31
3.2 Task Definition ..... 34
3.2.1 Stage 1 ..... 35
3.2.2 Stage 2 ..... 36
3.2.3 Stage 3 ..... 36
3.3 Multi-task Learning Model ..... 36



doi:10.6342/NTU202501600

3.3.1 Backbone model ..... 37
3.3.2 Task-specific Layer ..... 38
3.3.3 Loss Function ..... 38
Chapter 4 Experiment ..... 41
4.1 Dataset Summary Statistics ..... 41
4.1.1 Overall Summary Statistics ..... 41
4.1.2 Attribute Summary Statistics ..... 43
4.1.2.1 Tag ..... 43
4.1.2.2 Fact ..... 44
4.1.2.3 Time ..... 46
4.1.2.4 Measure ..... 47
4.1.2.5 Scale ..... 48
4.1.2.6 Decimals ..... 49
4.1.2.7 Axis and Member ..... 50
4.1.3 Textual Context Summary Statistics ..... 53
4.2 Experiment Implementation ..... 55
4.2.1 Multi-task Learning Model ..... 55
4.2.1.1 Data Preparation ..... 55
4.2.1.2 Model Architecture ..... 56
4.2.1.3 Loss Functions ..... 58
4.2.2 Few-shots Learning using LLMs ..... 58
4.2.3 Evaluation Metrics ..... 59
4.3 Experiment Result ..... 60
4.3.1 Result of Multi-task Learning Model ..... 60



doi:10.6342/NTU202501600

4.3.2 Result of Few-shot Learning ..... 61
4.4 Error Analysis ..... 63
4.4.1 Tag ..... 63
4.4.2 Time ..... 65
4.4.3 Scale ..... 67
4.4.4 Negative ..... 70
4.5 Managerial Implications ..... 71
Chapter 5 Conclusion ..... 74
REFERENCE ..... 76
APPENDIX ..... 84



doi:10.6342/NTU202501600

# LIST OF FIGURES


Figure 1 Data Demonstration ..... 4

Figure 2 The XBRL Instance Document Example ..... 8

Figure 3 An Example of Concept Definition in US-GAAP Taxonomy. ..... 9

Figure 4 The iXBRL File Example ..... 15

Figure 5 Example of iXBRL Element Tagging ..... 17

Figure 6 Example of Prior Dataset Limitation ..... 26

Figure 7 Sample Instance of Dataset ..... 32

Figure 8 The Simplified Time Attribute Tags ..... 35

Figure 9 Multi-task Learning Model Structure ..... 37

Figure 10 Positive and Negative Fact Value Distribution (Log Transform) ... 45

Figure 11 Numeric Value of Target Text Distribution (Log Transform) ..... 46

Figure 12 Context Length Distribution (Character Level and Token Level) .. 55

Figure 13 Tag Error Example ..... 65

Figure 14 Confusion Matrix of Time Prediction Result (Normalized) ..... 67

Figure 15 Confusion Matrix of Scale Prediction Result (Normalized) ..... 69

Figure 16 Scale Error Example ..... 70

Figure 17 Confusion Matrix of Negative Prediction Result (Normalized) ..... 71




doi:10.6342/NTU202501600

# LIST OF TABLES


| | | |
|---|---|---|
| Table 1 | Data structure and Description of Dataset | 33 |
| Table 2 | XBRL Target Attribute Description | 33 |
| Table 3 | Target Attribute of Three Stages | 34 |
| Table 4 | Summary Statistics by Year | 41 |
| Table 5 | Summary Statistics by Subset | 42 |
| Table 6 | Counts of Targets Contain Attribute | 42 |
| Table 7 | Unique Count of Attributes | 42 |
| Table 8 | Unique of Standard and Custom Tags | 43 |
| Table 9 | Counts of Standard and Custom Tags | 43 |
| Table 10 | Top-10 Frequent Tag Distribution | 44 |
| Table 11 | Unique and Counts of Frequent and Rare Tags in Training Set | 44 |
| Table 12 | Summary Statistics of Fact | 45 |
| Table 13 | Distribution of Negative and Non-negative Fact Value | 45 |
| Table 14 | Summary Statistics of Numeric Value of Target Text | 46 |
| Table 15 | Distribution of Time Category | 47 |
| Table 16 | Unique of Standard, Custom and Fraction Measures | 48 |
| Table 17 | Count of Standard, Custom and Fraction Measures | 48 |
| Table 18 | Distribution of Measure | 48 |
| Table 19 | Distribution of Scale | 49 |
| Table 20 | Distribution of Decimals | 50 |
| Table 21 | Number of Axis/Member per Target | 51 |
| Table 22 | Unique of Standard and Custom Axis | 51 |




doi:10.6342/NTU202501600

Table 23 Counts of Standard and Custom Axis ..... 51
Table 24 Unique of Standard and Custom Member ..... 52
Table 25 Counts of Standard and Custom Member ..... 52
Table 26 Top-10 Frequent Axis Distribution ..... 52
Table 27 Top-10 Frequent Member Distribution ..... 53
Table 28 Summary Statistics of Context Length ..... 54
Table 29 Hyper-parameter Setting ..... 57
Table 30 Result of Weighted Metrics ..... 60
Table 31 Result of Macro Metrics ..... 61
Table 32 Tag Hits@k ..... 61
Table 33 Result of Few-shot on LLMs ..... 62
Table 34 BLEU score and Jaccard Similarity ..... 63
Table 35 Most Frequent Tag Misclassified Class ..... 64
Table 36 Tag Class False Positive Rate ..... 65
Table 37 Time Misclassified Class ..... 66
Table 38 Time Class False Positive Rate ..... 66
Table 39 Most Frequent Scale Misclassified Class ..... 68
Table 40 Scale Class False Positive Rate ..... 69



doi:10.6342/NTU202501600

# Chapter 1 Introduction

Financial reports were often presented in various formats, such as PDF and HTML. While these formats were human-readable, their lack of standardization made automated data extraction and analysis difficult. Therefore, **eXtensible Business Reporting Language (XBRL)** was proposed. XBRL is an XML-based standard for exchanging business and financial information and is globally utilized by regulators, companies, and investors.

Specifically, XBRL enables tagging of financial data, allowing each data point to be associated with additional attributes, such as concepts (similar to line items in accounting, e.g., "Assets"), units, and reporting periods. XBRL reports are typically stored as separate XML files, which are machine-readable but not inherently user-friendly for human readers (XBRL International Inc, 2013; XBRL US, 2020). Subsequently, to achieve both human readability and machine readability within a single document, **inline XBRL (iXBRL)** was introduced. iXBRL embeds XBRL tags directly within an XHTML file, allowing financial reports to be presented in a human-friendly format while still maintaining a machine-readable structure for automated processing (Chang et al., 2021; SEC, 2018).

The growing adoption of XBRL and iXBRL in financial reporting has increased the need for automation to improve efficiency and consistency. Manual tagging is often labor-intensive and error-prone, especially in corporate settings where financial filings must meet strict regulatory standards. Automated iXBRL tagging reduces human effort, speeds up the reporting process, and enhances data reliability. In addition, models trained on tagged data can facilitate the understanding of other content that is not



doi:10.6342/NTU202501600

mandatorily the XBRL tagging, such as shareholder meeting materials or financial reports published before mandatory XBRL adoption. These benefits make automated tagging increasingly important in both research and industry (Khatuya et al., 2024; Sharma et al., 2023).

Recent efforts have contributed valuable work to support this task. FiNER-139 (Loukas et al., 2022) provides 1.1 million sentences annotated with 139 of the most frequent tags, while FNXL (Sharma et al., 2023) includes 79,088 sentences labeled with 2,794 tags. These datasets have advanced the development of models capable of identifying standard US-GAAP concepts from financial reports.

While these contributions are substantial, there remain opportunities for further enhancement. We identified several limitations in existing dataset. First, although XBRL consist of multiple attributes, prior works focus only on the tag name. Second, prior datasets include only US GAAP standard tags, potentially omitting tags from custom taxonomies or other non-US GAAP standards. Third, earlier datasets typically only provide the sentence that contains the tagged value, without including additional contextual information or document-level metadata that could improve tagging performance. Furthermore, we observed inconsistent and missing tags across identical sentences. Specifically, there are cases where the same sentence appears multiple times in the dataset, but only some instances are annotated with XBRL tags, while others are left untagged. Lastly, we also noticed noisy content in the dataset, such as "Table of Contents" or page numbers that do not contribute meaningful information for training.

To address these limitations, we construct a new dataset that: (1) expands the prediction scope to include multiple iXBRL attributes, such as time, scale, and measurement unit; (2) incorporates both standard and custom-defined taxonomy



doi:10.6342/NTU202501600

concepts; (3) enriches each instance with contextual and document-level metadata; (4) improves consistency by only extracting sentences that contain valid XBRL tags; and (5) utilizes iXBRL semantics mechanism to correctly handle cross-page tagging issues and filter out irrelevant content.

Our dataset comprises approximately 8,000 companies and over 1.1 hundred thousand filings from 10-K and 10-Q reports between years 2019 and 2024. It contains more than 6 million sentences and over 12 million target tags. The core task is defined as predicting specific iXBRL attributes for designated target sequences within sentences. Figure 1 demonstrates the example of data structure.

To improve the dataset's usability and support gradual model development, we structure the overall task into three stages, with this study focusing on the first stage. For baseline model, we adopt a multi-task learning framework, enabling simultaneous prediction of multiple attributes. Our experiments primarily leverage fine-tuned BERT-based models as the backbone architecture. In addition, we include few-shot prompting using Large Language Models (LLMs) as baseline comparisons to evaluate generalization capabilities without task-specific fine-tuning.

The rest of this thesis is structured as follows: Chapter 2 provides a detailed discussion of XBRL and iXBRL, related work of automated tagging in XBRL, application of large language models (LLM) in XBRL, and a review of multi-task learning. Chapter 3 presents the dataset construction process and dataset structure. We also introduce the multi-tasking learning methodology in detail. Chapter 4 reports the experimental implementation and results, while Chapter 5 concludes the study, discussing limitations and future research directions.



doi:10.6342/NTU202501600

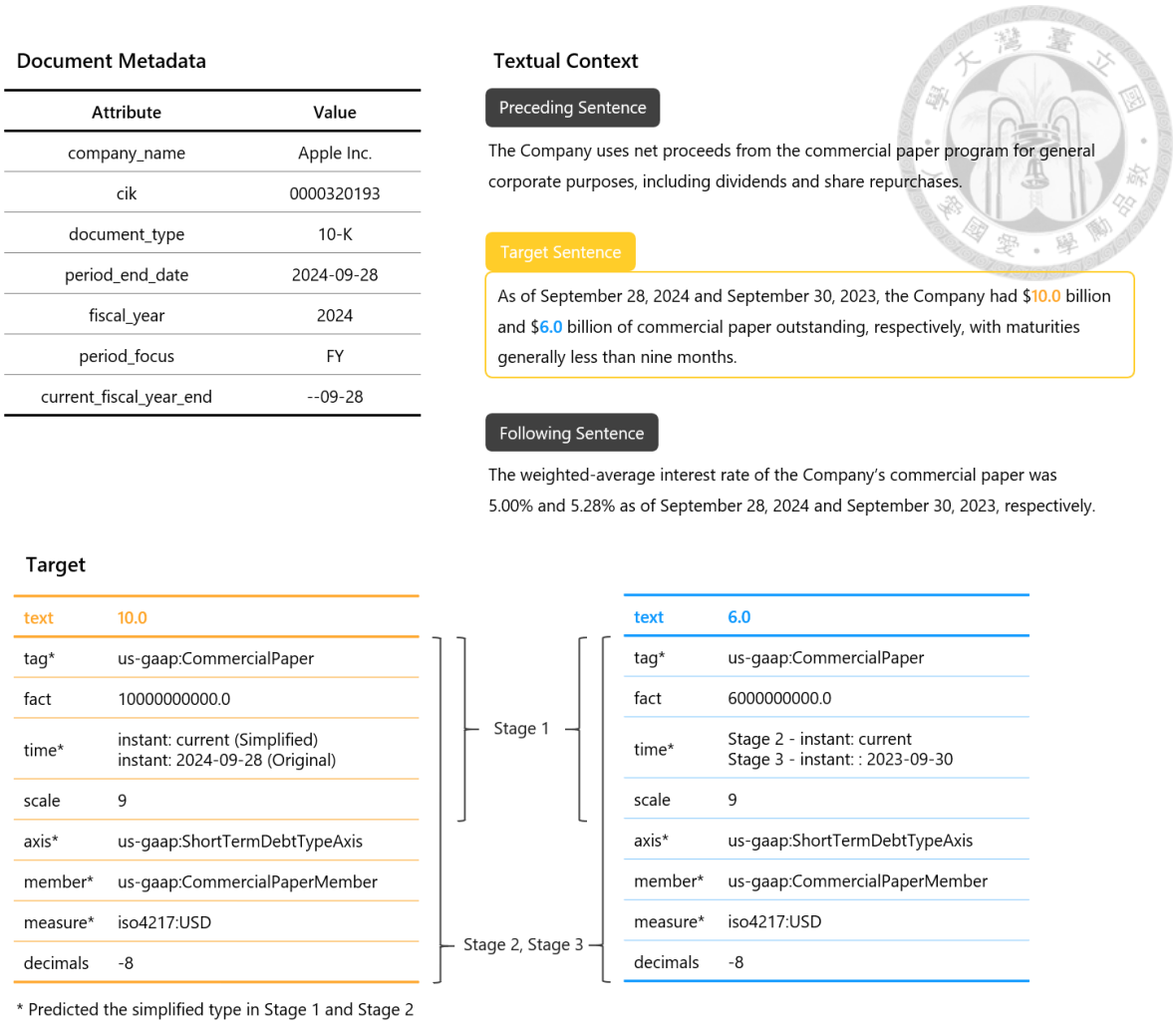



Figure 1 Data Demonstration



doi:10.6342/NTU202501600

# Chapter 2 Literature Review

In this chapter we provide an overview of XBRL and iXBRL, highlighting their significance in financial reporting and the challenges associated with their adoption. Then explores existing Natural Language Processing (NLP) approaches for processing XBRL data, including prior work on automated tagging and information extraction. Given that our study formulates XBRL attribute prediction as a multi-task learning problem, we further review relevant literature on multi-task learning, discussing common architectures, and optimization strategies.

## 2.1 XBRL

In this section, we present an introduction to XBRL, including the background, structure, and adoption of XBRL. Readers interested in in-depth discussions on these topics are referred to other excellent articles and books, including XBRL International,[1] XBRL 2.1 Specification (XBRL International Inc, 2013), XBRL Taxonomy Development Handbook (XBRL US, 2020), and The XBRL Book (Fourny, 2023).

### 2.1.1 Overview of XBRL

XBRL improves data interoperability and facilitates automation in financial reporting by embedding machine-readable data. Before the adoption of XBRL, financial reports were typically reported in various formats such as PDF, DOC, and HTML. While these formats are human-readable, their lack of standardization posed challenges for systematically collecting, analyzing, and comparing financial data.

[1] XBRL International: https://www.xbrl.org/



doi:10.6342/NTU202501600

Recognizing these limitations, a group of accountants and business experts, including members of the American Institute of Certified Public Accountants (AICPA), initiated the development of XBRL in 1998. This effort led to the establishment of XBRL International (XII), intending to create a standardized financial reporting language that enables seamless data exchange across different systems while preserving essential information. In December 2003, the first stable version of the XBRL specification was officially released, marking a significant milestone in the adoption of structured financial reporting (XBRL International, n.d.). Since then, XBRL has been widely adopted by regulators, companies, and investors to enhance financial transparency and reporting efficiency. The global status of XBRL adoption is provided in section 2.1.3.

### 2.1.2 XBRL Structure and Component

In the XBRL framework, taxonomies must comply with the XBRL specification, while instance documents must conform to both the taxonomy and the specification. This section provides an overview of the XBRL structure, including its basic components such as taxonomies, instance documents, reported facts, and concepts, along with their roles in ensuring standardized financial reporting.

#### 2.1.2.1 Reported Facts

XBRL enables data tagging, allowing software systems to recognize and process individual data items. Once tagged, each piece of information is represented as a *reported fact* in an XBRL report. A reported fact consists of a fact value along with its corresponding characteristics, which provide essential additional information for interpretation. These characteristics are commonly referred to as *dimensions* (or *aspects*). Dimensions can be broadly categorized into built-in dimensions and



doi:10.6342/NTU202501600

taxonomy-defined dimensions.

**Built-in dimensions** are required for all reported facts depending on their data type and are defined by the XBRL specification. These include:

- **Concept**: Provide the meaning for the reported fact, specified in the taxonomy (e.g., Assets, NetIncome).
- **Entity**: Refers to the reporting company or organization.
- **Period**: Specifies the time information to which the reported fact is reported.
- **Unit**: Indicates the unit of measurement (e.g., USD, shares).
- **Decimal**: Specifies the level of numerical precision. For example, a decimal value of 2 means the number is accurate to two decimal places (e.g., 1.2586 is rounded to 1.25), while a special value of INF denotes that the number is reported as exact with no rounding.

In contrast, **taxonomy-defined dimensions** are defined within an XBRL taxonomy and offer additional information to further classify reported facts. These include:

- **Axis**: A categorical grouping used to organize reported facts. (e.g., ProductLines).
- **Member**: A specific item within an axis that represents a distinct category.

### 2.1.2.2 XBRL Instance Document

An **XBRL instance document** is an XML-based file that contains a collection of reported facts along with the supporting metadata necessary for interpretation. To avoid redundant information, several dimensions are defined separately and referenced by multiple reported facts, enabling efficient data structuring. The key components of an instance document include:

- **Schema Reference**: Links to the taxonomy schema.
- **Contexts:** Defines reusable information, such as the reporting entity, reporting

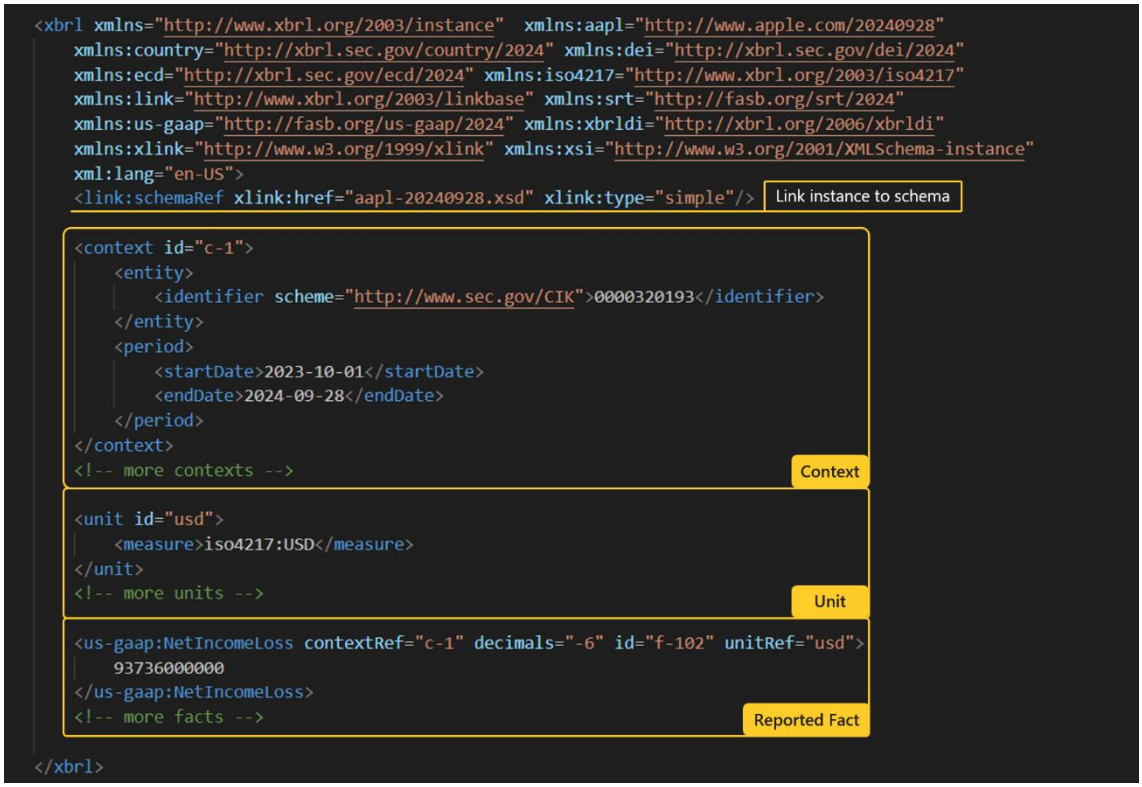


doi:10.6342/NTU202501600

period, axis, and member.

- **Units:** Specifies the measurement units for numerical reported facts.
- **Reported Facts:** Represent the actual financial data, each referencing a context and unit through their respective identifiers.

Figure 2 presents a part of the XBRL instance document from Apple Inc.'s 2024 10-K filing. It demonstrated a reported fact that represents a net income of 93,736 million USD for Apple Inc. from October 1, 2023, to September 28, 2024.




```
<xbrl xmlns="http://www.xbrl.org/2003/instance"  xmlns:aapl="http://www.apple.com/20240928"
    xmlns:country="http://xbrl.sec.gov/country/2024" xmlns:dei="http://xbrl.sec.gov/dei/2024"
    xmlns:ecd="http://xbrl.sec.gov/ecd/2024" xmlns:iso4217="http://www.xbrl.org/2003/iso4217"
    xmlns:link="http://www.xbrl.org/2003/linkbase" xmlns:srt="http://fasb.org/srt/2024"
    xmlns:us-gaap="http://fasb.org/us-gaap/2024" xmlns:xbrldi="http://xbrl.org/2006/xbrldi"
    xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xml:lang="en-US">
    <link:schemaRef xlink:href="aapl-20240928.xsd" xlink:type="simple"/>
    <context id="c-1">
        <entity>
            <identifier scheme="http://www.sec.gov/CIK">0000320193</identifier>
        </entity>
        <period>
            <startDate>2023-10-01</startDate>
            <endDate>2024-09-28</endDate>
        </period>
    </context>
    <!-- more contexts -->
    <unit id="usd">
        <measure>iso4217:USD</measure>
    </unit>
    <!-- more units -->
    <us-gaap:NetIncomeLoss contextRef="c-1" decimals="-6" id="f-102" unitRef="usd">
        93736000000
    </us-gaap:NetIncomeLoss>
    <!-- more facts -->
</xbrl>
```

Figure 2 The XBRL Instance Document Example

### 2.1.2.3 Concepts

A concept is a taxonomy-defined element that serves as the label and semantic identifier for a fact value in XBRL. It provides the necessary meaning for interpreting the reported data. Each concept is defined by several key attributes:

- **Qualified Name:** The unique identifier for the concept, consisting of a namespace



doi:10.6342/NTU202501600

and a local name.

- **Data Type:** Specifies the type of value the concept accepts (e.g., monetary, string, date).
- **Period Type:** Indicates whether the concept refers to a specific point in time (instant) or spans a duration (duration).
- **Balance Type:** Applicable to monetary concepts, indicating whether the value represents a credit or debit balance.
- **Nillable:** Determines whether the concept is allowed to have an empty value.
- **Abstract:** Indicates whether the concept is abstract. Abstract concepts are used for structural or organizational purposes within the taxonomy and cannot be directly used to tag reported facts in an instance document.

Figure 3 illustrates how a concept is defined in the US-GAAP taxonomy. It defines the NetIncomeLoss concept with monetary type, credit balance, and should be reported as duration.

```
<xs:element id='us-gaap_NetIncomeLoss' name='NetIncomeLoss'
    nillable='true' substitutionGroup='xbrli:item' type='xbrli:monetaryItemType'
    xbrli:balance='credit' xbrli:periodType='duration'
/>
```

Figure 3 An Example of Concept Definition in US-GAAP Taxonomy.

### 2.1.2.4 XBRL Taxonomy

**XBRL taxonomies** define concepts and relationships between those concepts that can be used in instance documents, thereby ensuring that XBRL reports are comparable. A taxonomy consists of two main components: the taxonomy schema file and a set of taxonomy linkbase files.

The **taxonomy schema** is an XSD file that defines individual concepts and their



doi:10.6342/NTU202501600

associated metadata, such as concept's name, data type, period type, and balance. In contrast, t**axonomy linkbases** are XML files that define various types of relationships among concepts. The primary types of linkbase include:

- **Label Linkbase:** Associates human-readable labels with machine-readable concept names, supporting multilingual representation.
- **Definition Linkbase:** Defines logical relationships between concepts, including taxonomy-defined dimensions such as axes and members.
- **Presentation Linkbase:** Organizes concepts into a hierarchical structure, which facilitates the structured rendering of financial statements.
- **Calculation Linkbase:** Specifies mathematical relationships (e.g., summation or subtraction) between numerical concepts, allowing XBRL processors to validate consistency in reported values.
- **Reference Linkbase:** Links concepts to external authoritative references, such as accounting standards or regulatory guidelines.

In addition, many regulatory bodies allow companies to develop **extended taxonomies** when the standard taxonomy does not fully accommodate their specific requirements. These extensions enable the definition of custom concepts or additional dimensions. While extensions offer flexibility, excessive customization may hinder comparability across different reporting entities.

### 2.1.3 XBRL Adoption

#### 2.1.3.1 Global Adoption of XBRL

Regulatory authorities in many jurisdictions have mandated XBRL-based filings, requiring publicly listed companies and financial institutions to submit their financial reports in this standardized format.



doi:10.6342/NTU202501600

China was among the earliest adopters, formally implementing XBRL reporting in 2004 (Liu, Luo, et al., 2014). South Korea mandated XBRL-based financial reporting for all publicly traded companies in 2007 (Yoon et al., 2011). Japan followed by requiring all filers to submit financial statements in XBRL format starting in April 2008 (FSA, 2008).

In the United States, the Securities and Exchange Commission (SEC) introduced XBRL reporting requirements in December 2008, beginning with a phased implementation in June 2009. By June 2011, XBRL reporting had become fully mandatory for all publicly listed companies (SEC, 2008). Similarly, Taiwan adopted XBRL in 2010, requiring listed companies to file their financial statements using the XBRL format (Taiwan Stock Exchange Corporation).

#### 2.1.3.2 Advantages of XBRL Adoption

The adoption of XBRL brings a wide range of benefits to various stakeholders, including investors, corporations, regulators, and auditors. Prior research has demonstrated that XBRL enhances the transparency and quality of financial disclosures, resulting in more reliable and informative financial reporting (Al-Okaily et al., 2024; Hodge et al., 2004; Liu, Wang, & Yao, 2014). Furthermore, Blankespoor (2019) found that firms tend to increase the level of footnote disclosure after adopting XBRL, likely due to reduced information processing costs for report users.

In addition to improving transparency, XBRL adoption enhances information efficiency and data usability through improves the overall presentation quality and searching efficiency and comparability of financial information (Chowdhuri et al., 2014; Tawiah & Borgi, 2022).

Moreover, XBRL adoption has been associated with positive impacts on capital



doi:10.6342/NTU202501600

markets. Several studies have identified a positive relationship between XBRL implementation and market liquidity, suggesting that XBRL reduces information asymmetry among market participants (Gupta et al., 2023; Liu et al., 2017; Yoon et al., 2011). Additionally, the use of XBRL has been linked to reductions in audit-related costs. For instance, Amin et al. (2018) reported a decreasing trend in audit report lag over time. It indicates enhanced reporting efficiency and audit quality, possibly due to a shortened learning curve as firms become more familiar with the XBRL framework.

#### 2.1.3.3 Challenge of XBRL Adoption

Despite the numerous advantages of XBRL, its implementation poses several challenges. A primary concern lies in the complexity of XBRL taxonomies. Accurate tagging of financial data requires specialized domain expertise, making the preparation process both time-consuming and technically demanding. As a result, the initial adoption of XBRL often involves high implementation costs and a steep learning curve for preparers (Chouhan & Goswami, 2015; Janvrin & No, 2012; Liu, Luo, et al., 2014)

Another major issue is the prevalence of errors in XBRL filings. Prior research has documented a wide range of common mistakes, such as missing elements, incorrect monetary values, and sign errors (Bartley et al., 2011; Du et al., 2013). Bartley et al. (2011) found that error rates were particularly high during the early stages of XBRL adoption. These errors declined significantly by the third year, attributed to increased preparer familiarity and improved processes. However, the continued presence of inaccuracies suggests persistent challenges in achieving filing precision and consistency.

## 2.2 Inline XBRL

In this section, we introduce inline XBRL (iXBRL), an advanced evolution of the



doi:10.6342/NTU202501600

XBRL format. We provide an overview of iXBRL, describe its key components and mechanisms, and discuss its adoption in regulatory environments. Most of the information presented in this section is based on XBRL International, Inline XBRL Specification 1.1 (XBRL International Inc., 2013), XBRL Book (Fourny, 2023), iXBRL Tagging Feature (XBRL International Inc., 2019).

### 2.2.1 Overview of iXBRL

With the adoption of XBRL, financial reports were typically submitted in two separate formats: an HTML version for human readability and an XBRL instance document for machine processing. However, XBRL instance documents often lack visual presentation and are not easily interpretable without specialized software. This has presented significant barriers to adoption, including increased complexity, a higher likelihood of tagging errors, and the need for technical expertise.

To address these limitations, iXBRL was introduced. iXBRL enables the embedding of XBRL tags directly into an HTML document, resulting in a single file that is both human-readable and machine-processable. This unified format eliminates the need to maintain separate HTML and XBRL files, thereby reducing redundancy and improving both the usability and accessibility of financial reports.

### 2.2.2 iXBRL Structure and Tagging

#### 2.2.2.1 iXBRL Report Structure

iXBRL is an XHTML-based format that embeds XBRL data into standard HTML documents, enabling seamless integration of presentation and structured data extraction. In this format, standard HTML elements (e.g., , , <table>) define the visual layout of the report and can rendered directly in web browsers without need for



doi:10.6342/NTU202501600

additional software. Meanwhile, iXBRL elements ensure that the underlying structure data remains accessible to XBRL processors. Figure 4 provides an example of an iXBRL document from Apple Inc.'s 2024 10-K Filing. Several key iXBRL elements include:

- <ix:header>: A child node of a non-display element. It contains essential metadata required for the interpretation and validation of the report:
  - <ix:hidden>: Stores instance data that is not visible in the browser but is crucial for XBRL validation.
  - <ix:references>: Specifies references to the taxonomy schemas.
  - <ix:resources>: Includes <xbrli:context> and <xbrli:unit> elements, which define contextual information such as reporting periods, entity identifiers, and measurement units.
- The financial statement information is rendered on the browser and includes mainly information about the filing such as the reported fact, footnotes, textual discourse, etc. The iXBRL element used for tagging reported facts includes:
  - <ix:nonFraction>: Used to tag numeric reported facts, such as monetary values, integers, and share counts.
  - <ix:nonNumeric>: Used to tag non-numeric reported facts, including textual content and date values.



doi:10.6342/NTU202501600


Figure 4 The iXBRL File Example

### 2.2.2.2 iXBRL Tagging

iXBRL modifies the presentation of XBRL reports without altering the amount of disclosed information (Chang et al., 2021). In an iXBRL document, reported facts are embedded directly within the XHTML content using specific elements such as <ix:nonFraction> for numerical values and <ix:nonNumeric> for textual information, dates, and other non-numeric data. These tagged reported facts maintain attributes consistent with those in standard XBRL.

To improve both data readability and usability, iXBRL introduces several mechanisms, including support for format, scale, and sign attributes, along with advanced features like continuation and exclusion for handling complex textual content.



doi:10.6342/NTU202501600

#### 2.2.2.2.1 Format

Certain data types are presented in a more human-readable format in iXBRL, such as numbers with thousand separators or dates in a localized format. To ensure accurate data extraction, the format attribute is used to defined how these values should be interpreted. For instance, a date displayed as "September 28, 2024" in an iXBRL report would use the datemonthdayyearen format, corresponding to the standardized XBRL value "2024-09-28." Similarly, a numerical value such as "123,000.21" would be tagged with the numdotdecimal format, indicating that a dot is used as the decimal separator and a comma may be used for thousands grouping.

#### 2.2.2.2.2 Scale

To present numerical values more concisely, iXBRL supports a scaling mechanism via scale attribute. This attribute specifies a power of ten by which the displayed number should be multiplied to recover its unscaled XBRL value.

For example, a reported value of 210 in millions would include a scale="6" attribute, indicating it should be multiplied by $10^6$ to yield 210,000,000. Conversely, for percentages such as 19%, a scale="-2" would be applied, meaning the number should be multiplied by $10^{-2}$ to convert it to the machine-readable value 0.19.

#### 2.2.2.2.3 Sign

iXBRL also offers mechanisms to ensure the correct interpretation of the sign of numerical values. When a value is explicitly negative, it is indicated using the sign="-" attribute. For example, if a reported fact tagged with us-gaap:NetIncomeLoss represents a net loss, the sign attribute would be set to '-'.

However, it is important to note that taxonomy concepts are generally intended to



doi:10.6342/NTU202501600

be reported as positive values. As such, the sign attribute is often omitted, and values are normalized to positive numbers when extracted as XBRL data.

Negative values may still be visually presented in the human-readable report using formatting conventions, such as prefixing the value with a minus sign or enclosing it in parentheses, without affecting the underlying machine-readable data. For example, "Cost of Goods Sold (COGS)" may appear as a negative figure on the income statement for clarity, but is reported positively in XBRL, with the visual cue added through HTML styling or formatting. Figure 5 demonstrates an example integrating the format, scale, and sign attributes for a numeric reported fact within an iXBRL report.

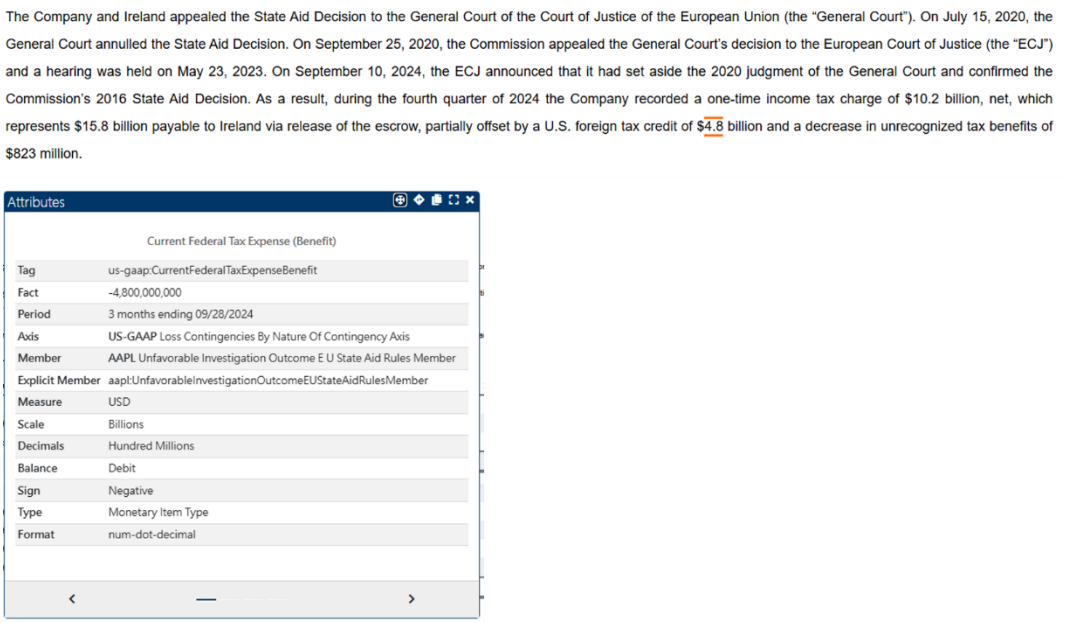



Figure 5 Example of iXBRL Element Tagging

## 2.2.2.2.4 Textual data

In iXBRL, the tagging of textual content, which is often found in footnotes and management discussion, plays a vital role in ensuring that narrative disclosures are accurately structured for machine processing. To accommodate complex textual information, iXBRL introduces two key mechanisms: continuation and exclude.



doi:10.6342/NTU202501600

The <ix:continuation> element allows a single reported fact to span multiple, non-contiguous portions of the XHTML document. This is particularly useful for long passages of text that are split across different sections or pages. On the other hand, the <ix:exclude> element is used to omit specific parts of the content from XBRL extraction, without altering the visual or structural integrity of the document. This feature ensures that extraneous or decorative text (e.g., labels, footers) does not interfere with the accuracy of extracted values.

### 2.2.3 iXBRL Adoption

#### 2.2.3.1 Global Adoption of iXBRL

The adoption of iXBRL has progressed globally, with numerous regulatory authorities mandating its use for financial and tax reporting. The United Kingdom, where iXBRL technology originated, was the first country to implement its use. Since 2011, the HM Revenue & Customs (HMRC) has required companies to submit tax filings in iXBRL format (HMRC, 2020).

In Japan, the FSA adopted iXBRL in 2014, mandating its use for financial statement submissions (FSA, 2013). Similarly, the SEC in the United States, the introduced a phased implementation plan for iXBRL reporting through amendments passed in June 2018. The implementation began in 2019, with a full transition requiring all filers to submit reports in iXBRL by 2021 (SEC, 2018). In Taiwan, regulatory authorities requiring financial reports to be submitted in iXBRL began in the first quarter of 2019 (Taiwan Stock Exchange Corporation). The European Securities and Markets Authority (ESMA) also mandated the use of iXBRL within the European Single Electronic Format (ESEF). Approved in December 2018, this regulation required companies to submit financial reports in XHTML format starting in 2020. Furthermore,



doi:10.6342/NTU202501600

companies preparing IFRS consolidated financial statements must embed XBRL-tagged data within the XHTML format using iXBRL. (ESMA)

#### 2.2.3.2 Advantages of iXBRL Adoption

The adoption of iXBRL brings several benefits to financial reporting. One of the primary advantages is its ability to enhance the usability, accessibility, and transparency of disclosures by combining human-readable content and machine-readable data (Basoglu & White, 2015). This dual readability allows users to navigate and analyze structured financial information more efficiently, aiding decision-making processes for both professional analysts and individual investors (Chang et al., 2021; Zhang & Shan, 2024). Additionally, it encourages broader engagement from individual investors, ultimately contributing to long-term reductions in information asymmetry (Luo et al., 2023).

For preparers, iXBRL offers increased control over how financial reported facts are presented within reports. It enables more flexible formatting and eliminates the need to maintain separate HTML and XBRL versions (SEC, 2018). Moreover, iXBRL helps reduce tagging inconsistencies and common errors, thereby improving the overall quality of reported data (Basoglu & White, 2015)

From a cost-efficiency perspective, the SEC (2018) has noted that iXBRL reduces duplicative efforts, simplifies validation procedures, and may lower long-term compliance costs. Furthermore, research suggests that iXBRL contributes to improved lower capital costs in subsequent reporting periods by enhancing communication efficiency and reducing information asymmetry (Luo et al., 2023).



doi:10.6342/NTU202501600

## 2.3 NLP, LLM, and XBRL

This section introduces recent applications and research involving the integration of NLP and LLM with XBRL. We first examine how LLMs are utilized to analyze and interpret XBRL data, followed by an overview of prior work aimed at automating the XBRL tagging process.

### 2.3.1 LLM-based Analysis of XBRL

Recent studies and articles have explored the application of large language models (LLMs) in the analysis of XBRL-based financial reports. Ramanan (2024a), XII Guidance Manager, explored the use of GPT-4 (OpenAI et al., 2023) in understanding and interpreting XBRL data, particularly through the XBRL-JSON format (a transformed format of inline XBRL). The articles demonstrated that GPT-4 can effectively respond to user queries, such as retrieving the percentage of profit growth in the previous year. However, the author emphasized that accurate interpretation depends on the model's ability to identify the correct concept or dimension, and pointed out ongoing challenges related to consistency and reliability.

Beyond basic question answering, Ramanan (2024b, 2025) also proposed a broader range of use cases for GPT-4 in financial reporting tasks. These include analyzing narrative disclosures to detect changes in accounting policies or assess potential greenwashing by classifying the objective and subjective disclosures. Furthermore, LLMs were discussed as tools to support taxonomy improvements, such as enhancing label clarity, promoting consistency in definitions, and simplifying the interpretation of calculation and formula relationships.

In addition, Han et al. (2024) proposed the XBRL Agent framework, which



doi:10.6342/NTU202501600

integrates external tools to enhance LLM performance. It incorporates a retriever based on Retrieval-Augmented Generation (RAG) to access the latest financial information, along with a calculator module to support accurate numerical computation. Experimental results on tasks involving financial domain queries and numeric calculations show that these enhancements improve the accuracy and reliability of LLMs in financial report analysis.

### 2.3.2 Automated Tagging in XBRL

There is growing interest in leveraging NLP and LLM techniques to enhance the efficiency and accuracy of the tagging process. XBRL International Inc (2024) has highlighted the potential of artificial intelligence (AI) to streamline digital tagging and reduce overall tagging costs. These tools can automate repetitive tagging tasks while allowing human experts to focus on complex cases requiring domain knowledge.

A report by UBPartner's (2024), an XBRL software vendor, applied NLP to automatically assign tags in tabular financial data. The method was effective for structured formats like balance sheets and cash flow statements but faced limitations when dealing with complex or irregular table structures.

FiNER-139 (Loukas et al., 2022) is a dataset that formulates tagging as a named entity recognition (NER) problem, focusing on 139 commonly used US-GAAP elements. This dataset includes approximately 1.1 million sentences from 10,000 financial reports (2016–2020), and the authors proposed a BERT-based (Devlin et al., 2019) model pre-trained on financial texts with a masking strategy to handle fragmented numerical expressions.

FNXL (Sharma et al., 2023) expanded the tagging scope to 2,794 US-GAAP tags using 79,000 sentences extracted from 10-K filings by 2,239 companies (2019–2021).



doi:10.6342/NTU202501600

Their approach involved a two-stage model: (1) a binary BERT-based tagger to detect relevant values and (2) AttentionXML (You et al., 2019) Model, an extreme classification model, to assign labels.

FLAN-FinXC (Khatuya et al., 2024) further advanced tagging performance on the FiNER-139 and FNXL datasets by instruction-tuning FLAN-T5 (Chung et al., 2024) and using the parameter-efficient LoRA (Hu et al., 2022) technique. Their two-stage generative framework first uses a large language model to generate XBRL tag documents (human-readable descriptions for concepts) if the numeral is relevant. In the second stage, then applies a sentence encoder, Sentence-T5-XXL (Ni et al., 2022), computes embeddings for both the predicted XBRL tag document from the first stage and the ground truth tag documentation then selects the tag with the highest cosine similarity as predicted tags.

## 2.4 Multi-task Learning

Multi-task learning (MTL) aims to improve generalization across multiple related tasks by enabling knowledge sharing between them (Caruana, 1997; Zhang & Yang, 2022). Compared to single-task learning, MTL has been shown to enhance model performance, foster more robust representations, and reduce the risk of overfitting (Zhang & Yang, 2022). MTL is implemented by optimizing multiple loss functions simultaneously (Ruder, 2017).

According to Ruder (2017), deep learning-based MTL architectures can be categorized into two types. The first is hard parameter sharing, where lower-layer representations are shared across tasks, while task-specific layers are maintained at higher levels. This approach reduce overfitting by encouraging the model to learn



doi:10.6342/NTU202501600

generalizable features. The second is soft parameter sharing, where each task is assigned its own model, and regularization constraints are introduced to promote similarity between model parameters. This method enables loosely related tasks to share information while preserving flexibility in learning task-specific representations.

Zhang et al. (2023) further classify MTL approaches in NLP based on task relatedness, distinguishing between joint training and multi-step training. Joint training is used when tasks can be learned simultaneously, typically through either hard or soft parameter sharing. In contrast, multi-step learning is applied when a task's input depends on the output or hidden states of preceding tasks, forming a sequential learning pathway through task-specific decoders.

A common optimization strategy in MTL involves minimizing a weighted sum of loss functions from different tasks and updating model parameters via gradient descent The loss weights can be either predefined or dynamically adjusted (Zhang et al., 2023).

Joint training in text classification involves two common settings: (1) a single input with multiple outputs, where several attributes are predicted simultaneously, and (2) multiple inputs with multiple outputs, where tasks are trained in parallel with distinct inputs. For instance, Yu et al. (2018) leverage sentiment classification to improve emotion classification, while Gui et al. (2022) combine sentiment classification and topic detection in a MTL framework to enhance the performance of both tasks.

## 2.5 Research Gap

Prior works on automated XBRL tagging have laid an important foundation, particularly in identifying standard US-GAAP concept names within financial texts. However, there is still considerable room for improvement in terms of contextual



doi:10.6342/NTU202501600

completeness and practical applicability of the tagging system.

Existing datasets typically focus only on predicting concept names, restrict tags to US-GAAP taxonomy, and provide only the tagged sentence, without including surrounding context or document-level metadata that could aid real-world tagging performance.

In addition, we observed missing and inconsistent tag within the same sentence across different instances. For example, in the training set of FiNER-139, we identified 163,869 groups of duplicated sentences, among which 895 groups (2,778 data points) contain instances where some sentences lack the expected XBRL tags. This may arise from duplicated descriptions appearing in various parts of financial report or incomplete tagging in earlier filing years, where some occurrences are left untagged. Moreover, the dataset also contains noisy content, such as page numbers and metadata text like "Table of Contents" embedded in the sentence body. Examples 1 to 5 in Figure 6 illustrate the issue of inconsistent or missing XBRL tags, while Examples 6 to 8 demonstrate the presence of noisy content within sentences.

More specifically, in Example 1, both instance IDs 830442 and 248100 contain the same sentence: "*Total Leverage Ratio "has the meaning given it in the facility . Our obligations under our second lien term loan facility are guaranteed by most of our wholly - owned U.S. and Canadian subsidiaries and are secured by second priority security interests in the same collateral securing the $ 2.0 billion first lien revolving credit facility* ." However, only ID 830442 includes an XBRL tag which the token “2.0” is tagged with B-LineOfCreditFacilityMaximumBorrowingCapacity. In contrast, ID 248100 does not include any XBRL tags; all tokens are labeled as O (outside).

On the other hand, Example 6 presents a sentence containing irrelevant embedded



doi:10.6342/NTU202501600

content: "*As a result , the entity which previously owned the land remains liable to the Federal Government tenant for the completion of the construction obligations* ***35 Table of Content*** *under the lease* ." The phrase "35 Table of Content" is irrelevant. It is likely a result of cross-page text extraction, where "35" refers to a page number and "Table of Content" appears at the top of the page in the original filing.

In Example 7, aside from page numbers and "Table of Content," the remaining content appears to be a heading from a note section, rather than a complete or semantically meaningful sentence.



doi:10.6342/NTU202501600

**Example 1**

| | |
|---|---|
| ID / Sentence | Total Leverage Ratio " has the meaning given it in the facility . Our obligations under our second lien term loan facility are guaranteed by most of our wholly - owned U.S. and Canadian subsidiaries and are secured by second priority security interests in the same collateral securing the $ 2.0 billion first lien revolving credit facility . |
| 830442 | B-LineOfCreditFacilityMaximumBorrowingCapacity |
| 248100 | (No tags) |

**Example 2**

| | |
|---|---|
| ID / Sentence | $ 1 billion Aggregate Principal Amount of 4.75 % Senior Notes due June 2023 ( the " 2023 Notes " ) . |
| 893172 | B-DebtInstrumentFaceAmount B-DebtInstrumentInterestRateStatedPercentage |
| 808454 | (No tags) |

**Example 3**

| | |
|---|---|
| ID / Sentence | ( 4 ) Beginning on August 8 , 2019 and at any time thereafter , the notes are subject to redemption at the Company 's option , in whole but not in part , at a redemption price equal to 100 % of the principal amount of the notes , plus unpaid accrued interest thereon to the redemption date . |
| 651146 | B-DebtInstrumentRedemptionPricePercentage |
| 432404 | (No tags) |

**Example 4**

| | |
|---|---|
| ID / Sentence | ( b ) The company has an 8.4 % equity ownership interest in Marubun Corporation and a portfolio of mutual funds with quoted market prices . |
| 854152 | B-EquityMethodInvestmentOwnershipPercentage |
| 180548 | (No tags) |

**Example 5**

| | |
|---|---|
| ID / Sentence | 13 Table of Contents The calculation of the weighted average number of shares of common stock outstanding for Basic EPS and Diluted EPS for the periods indicated ( in thousands , except per share data ) is as follows : Note 5 - Segment Reporting Prior to the Company ' s sale of Express Jet on January 22 , 2019 , the Company ' s three reporting segments consisted of the operations of Sky West Airlines , Express Jet and Sky West Leasing activities . |
| 627007 | B-NumberOfOperatingSegments |
| 347849 | (No tags) |

**Example 6**

| | |
|---|---|
| ID / Sentence | As a result , the entity which previously owned the land remains liable to the Federal Government tenant for the completion of the construction obligations 35 Table of Content under the lease . |
| 721864 | (No tags) |

**Example 7**

| | |
|---|---|
| ID / Sentence | 16 Table of Content BOSTON PROPERTIES , INC . AND BOSTON PROPERTIES LIMITED PARTNERSHIP NOTES TO THE CONSOLIDATED FINANCIAL STATEMENTS1 . |
| 401692 | (No tags) |

**Example 8**

| | |
|---|---|
| ID / Sentence | For land leases classified as finance leases because of a purchase option that the Company views as an economic incentive , the Company follows its existing policy and does not depreciate land because it is assumed to 26 Table of Content have an indefinite life . |
| 714351 | (No tags) |

Figure 6 Example of Prior Dataset Limitation



doi:10.6342/NTU202501600

# Chapter 3 Methodology

In this chapter, we present the methodology of our study, including the dataset construction process, task definition, and the design of multi-task learning model.

Note that in XBRL documentation, the term “fact” typically refers to a unit of reported information that combines a concept, value, context, and unit. To avoid confusion in this thesis, we distinguish between two related terms. We use “reported fact” to refer to the full information unit as defined in XBRL, whereas “fact” in our dataset and analysis specifically denotes the fact value, which is the numerical or textual content of the reported fact.

## 3.1 Dataset

To overcome the limitations identified in prior studies, we developed a new dataset. This section describes how our dataset is designed to bridge those gaps and provides the details of the data construction process and the resulting dataset structure.

### 3.1.1 Dataset Overview

We constructed a new dataset to address the limitations mentioned above. Our dataset expands the prediction scope beyond the tag attribute to include additional XBRL attributes including time, fact, scale, decimals, axis, and member. It also incorporates both standard and custom-defined taxonomy concepts, enabling broader applicability in real-world use cases. Furthermore, each instance is enriched with contextual and document-level information to offer a more comprehensive view, such as the preceding and following sentences, reporting entity, and reporting period end date.

To mitigate the issue of missing tags and noisy content, each instance in our dataset



doi:10.6342/NTU202501600

contains at least one valid XBRL tag. In addition, we leverage the ix:continuation tag in the iXBRL, to reconstruct cross-page text spans, ensuring that the extracted textual content is both complete and meaningful for the tagging task.

Our dataset is constructed from iXBRL-format 10-K and 10-Q filings collected from the SEC's EDGAR (Electronic Data Gathering, Analysis, and Retrieval) system, encompassing reports from approximately 8,000 companies filed between January 2019 and November 2024. In total, the dataset comprises around 6.6 million sentences extracted from 110,000 filings. We specifically chose iXBRL as the source format because it enables simultaneous access to both structured XBRL tags and their corresponding human-readable text. This integration not only aligns with recent SEC mandates but also makes it more suitable for developing dataset for XBRL attribute prediction. Similar to prior work (Loukas et al., 2022; Sharma et al., 2023), we focus on numeric tags appearing in narrative paragraphs, excluding those in tabular formats.

### 3.1.2 Dataset Construction Process

The dataset construction pipeline was designed to extract structured information from iXBRL files and convert it into a JSON/JSONL format. The process consists of the following steps:

**Step 1: Company List Extraction**

We extracted a list of companies that submitted 10-K and 10-Q reports between 2019 and 2024 using the SEC's directory listings of full index[2] files.

**Step 2: Recent Filing Retrieval**

For each listed company, we retrieved filing information from the SEC-provided

[2] EDGAR full index: https://www.sec.gov/Archives/edgar/full-index/



doi:10.6342/NTU202501600

API,[3] including the accession number (a unique identifier for each filing), the name of the primary iXBRL document, and the report date.

**Step 3: iXBRL File Download**

Based on the retrieved filing information, we generated URLs[4] for the iXBRL files and downloaded them using Python's requests library for further processing.

**Step 4: File Parsing and Processing**

For each year and company, the downloaded iXBRL files were parsed using BeautifulSoup library. The parsing process included the following components:

- **Document Metadata Extraction**

  Identified metadata tags, primarily those with the dei (Document and Entity Information) prefix. They include details such as the reporting date and company name. A full list of collected tags is provided in the Appendix.

- **Contextual Information Extraction**

  Extracted <xbrli:context> and <xbrli:unit> elements to build contextRef and unitRef dictionaries, which are essential for determining attributes like period, axis, and units of measurement. If the unit is of the fraction type, it is converted into the {numerator}/{denominator}. For instance, iso4217:USD/xbrli:shares represents U.S. dollars per share.

- **Textual Content Extraction**

  First, we removed tables and table-like structures from iXBRL document. Next, we extracted all <ix:nonfraction> elements, which represent tagged numerical values.

---

[3] EDGAR APIs: https://www.sec.gov/search-filings/edgar-application-programming-interfaces

[4] iXBRL link: https://www.sec.gov/Archives/edgar/data/{cik}/{accession_number}/{primary_document}



doi:10.6342/NTU202501600

To address the missing tag issue found in prior work, we reversed the pipeline by starting from the tags themselves, rather than parsing the HTML with regular expressions. This ensures that each selected sentence indeed contains a valid XBRL tag. For each element, we performed the following step:

1. **Paragraph Identification:** Locate the paragraph containing the target element, resolving cross-page issues to ensure the paragraph remains complete and without unnecessary context such as page number. To address the noisy content problem (e.g., page numbers or title), we leveraged the <ix:continuation> elements to concatenate content spread across pages. This helps preserve the semantic integrity and continuity of the paragraph.
2. **Tag Position Encoding:** Calculate the character-level offset of the target tag within the paragraph by matching it with surrounding words.
3. **Sentence Segmentation:** Applied the spaCy library to segment the paragraph into individual sentences and record their offsets.
4. **Textual Context Capture:** Identified the preceding and following sentences of the target sentence to provide local context.
5. **Attribute Extraction:** Determine the tag's position within the sentence based on its offset in the paragraph and extract the following attributes:
    - **Fact:** Computed using the value, scale, decimals, and sign.
    - **Tag, Scale, Decimals:** Directly parsed from element attributes.
    - **Period**, **Axis**, **Member**, and **Unit**: Derived from contextRef and unitRef mappings.

**Step5: Data Formatting**

The extracted information was organized into a structured JSON format.



doi:10.6342/NTU202501600

**Step6: Post-processing**

We applied heuristic rules to clean the context and document metadata. Tag positions were re-validated and adjusted to resolve inconsistencies caused by whitespace and HTML markup. Finally, data from all years were merged and split into training (79%), validation (10%), and test (11%) sets. To improve loading efficiency, these subsets were stored in JSONL format.

In addition, we organized the dataset into multiple stages with increasing levels of difficulty to support progressive evaluation and training. These task stages are described in detail in Section 3.2, and the dataset was formatted accordingly for each stage.

### 3.1.3 Data Structure

Each data instance consists of four main components: document, context, targets, and golds. The **document** component contains essential metadata about the filing, including the company name, Central Index Key (CIK), document type (e.g., "10-K" or "10-Q"), reporting period end date, fiscal year, and a direct URL to the corresponding iXBRL file hosted on the SEC's EDGAR system.

The **context** component provides textual surroundings for each target sequence. It includes the target sentence containing the numerical value to be tagged, along with its preceding and following sentences.

The **targets** component lists the target sequence in the target sentence that requires XBRL tagging. Each target sequence is assigned a unique sequence ID and annotated with its character-level start and end positions in the sentence. A set of attributes is associated with each target. Note that some attributes may be missing, depending on the availability of tagging information in the source document.

The **golds** component provides the ground-truth annotations for the XBRL target



doi:10.6342/NTU202501600

attributes associated with each target. Attribute values appear in the same order as the corresponding targets to ensure alignment.

Figure 7 demonstrate an example instance of our dataset. A summary of key fields in each component is provided in Table 1, while detailed descriptions of XBRL target attribute are presented in Table 2.


Figure 7 Sample Instance of Dataset



doi:10.6342/NTU202501600

Table 1 Data structure and Description of Dataset

| Component | Key | Description |
|---|---|---|
| job_id | - | Unique identifier for the instance. |
| document | company_name | Name of the company. |
| | cik | Central Index Key. |
| | document_type | Filing type (e.g., "10-K", "10-Q"). |
| | period_end_date | End date of the reporting period. |
| | fiscal_year | Fiscal year the document pertains to. |
| | document_link | URL to the original iXBRL file. |
| context | context_t | Target sentence containing the numerical value. |
| | context_p | Preceding sentence. |
| | context_n | Following sentence. |
| targets (list) | seq_id | Unique identifier for the target sequence. |
| | start_pos | Character-level start positions of the target span. |
| | end_pos | Character-level end positions of the target span. |
| | text | The actual word sequence requiring tagging. |
| | attributes | List of XBRL attributes to predict. |
| golds (list) | seq_id | Matches the identifier from the targets section. |
| | value | Ground truth values for the corresponding attributes. |

Table 2 XBRL Target Attribute Description

| Attribute | Description |
|---|---|
| tag | The taxonomy element's standard or custom identifier. |
| fact | The factual value, represented as a float. |
| time | The reporting period, expressed as:<br>• For a period: "start: yyyy-mm-dd; end: yyyy-mm-dd".<br>• For an instant: "instant: yyyy-mm-dd".<br>In stage 1 and stage 2, the specific date is transformed into a simplified format, such as past, current, and future. |



doi:10.6342/NTU202501600

| | |
|---|---|
| measure | The unit of measurement. |
| decimals | The decimal precision levels. |
| scale | The scaling factors. |
| axis (list) | A list of dimensions classifying the fact. Note that order of the axis is not significant.<br>Each axis item contains:<br>• amseq: Matches the identifier to the member.<br>• axis: The axis name. |
| member (list) | A list of specific members within an axis, each member item contains:<br>• amseq: Matches the identifier from the axis.<br>• member: The member name. |

## 3.2 Task Definition

The task objective is to predict specific XBRL attributes for target text sequences within sentences in a financial report. To gradually improve the model's ability to predict XBRL attributes and enhance the dataset's usability, we divided our task into three stages based on the number of attributes to be predicted and the complexity of those attributes. The target attribute of three stage is as shown in following Table 3.

Table 3 Target Attribute of Three Stages

| Attribute | Stage 1 | Stage 2 | Stage 3 |
|---|---|---|---|
| tag | ▲ | ▲ | ✓ |
| fact | ✓ | ✓ | ✓ |
| time | ▲ | ▲ | ✓ |
| axis | - | ▲ | ✓ |
| member | - | ▲ | ✓ |
| measure | - | ▲ | ✓ |
| decimals | - | ✓ | ✓ |
| scale | ✓ | ✓ | ✓ |

✓: Predicted the original value of the attribute ▲: Predicted the simplified type of the attribute



doi:10.6342/NTU202501600

### 3.2.1 Stage 1

In Stage 1, the target attributes include tag, fact, time, and scale, with certain modifications applied to specific attributes:

- Tag: Custom tags are labeled as custom, while standard tags occurring fewer than 1,000 times in the training set are labeled as standard_rare, following the threshold used in FiNER-139. Standard tags appearing more than 1,000 times in training set retain their original names. In addition to us-gaap, other namespaces that considered to be standard is provided in Appendix, such as dei and srt.
- Time: As illustrated in Figure 8, the time attribute is categorized based on two period types: instant and period. Depending on the period type, time is further classified into past, current, future, or combinations thereof (e.g., past_current, current_future). The accounting period varies depending on the document type. Specifically, 10-K filings correspond to a one-year reporting period, while 10-Q filings correspond to a three-month reporting period.

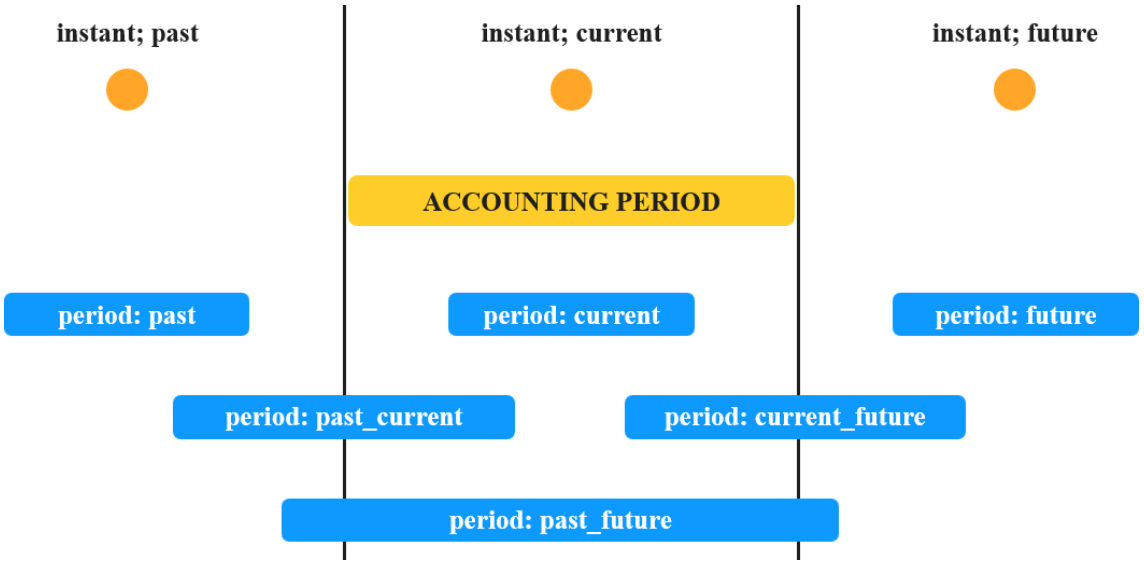



Figure 8 The Simplified Time Attribute Tags



doi:10.6342/NTU202501600

### 3.2.2 Stage 2

In Stage 2, additional attributes are introduced beyond those in Stage 1, which include tag, fact, time, and scale. Specifically, Stage 2 incorporates the axis, member, measure, and decimals attributes. For simplification, only attributes from standard namespaces are retained for axis, member, and measure, while custom-defined values are uniformly labeled as custom. The time attribute is simplified as Stage 1.

### 3.2.3 Stage 3

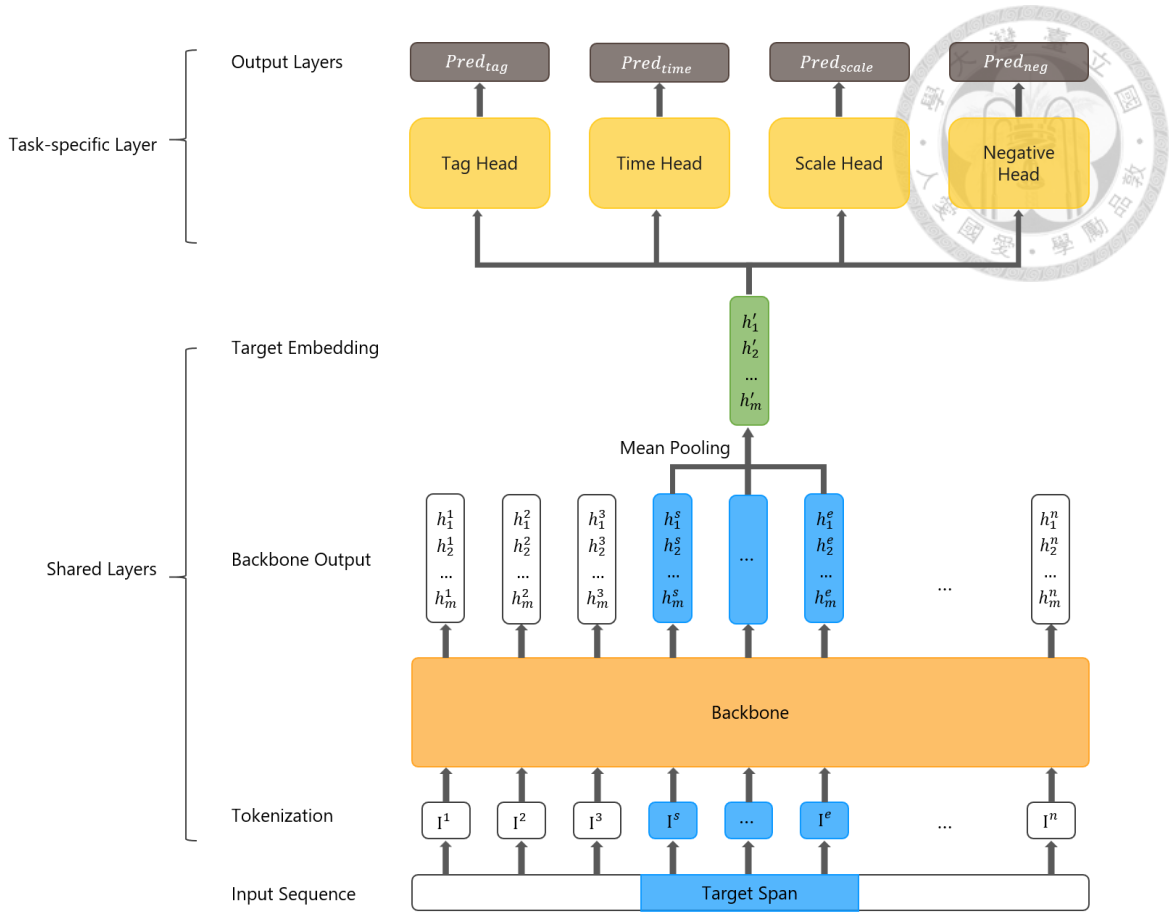
Stage 3 retains the same target attributes as Stage 2, but introduces increased task complexity. For tag, axis, member, and measure, the model is required to generate the actual custom content rather than simply predicting a generic custom label. Additionally, for the time attribute, the task shifts from predicting categorized time periods to identifying the exact reporting dates.

## 3.3 Multi-task Learning Model

We adopt multi-task learning (MTL) model to predict attributes simultaneously. Several backbone models serve as shared layers, where parameters are jointly optimized across different attribute prediction tasks. For attribute prediction tasks, task-specific output head consisting of a feedforward neural network. The model structure is illustrated in Figure 9.



doi:10.6342/NTU202501600




$I^i$: Tokenized input token at position i, where i ∈ {1, … , n}, and n is the length of the token sequence.
s, e: Start and end token positions of the target span.
$h^i_j$: Last hidden state output from the backbone model at token position i and hidden dimension j, where j ∈ {1, … , m}, and m is the hidden size.
$h'_j$: Mean pooled hidden representation of the span from position s to e, computed as the average of hidden states $h^i$ for i ∈ [s, e].

Figure 9 Multi-task Learning Model Structure

### 3.3.1 Backbone model

The backbone models employed are described as follows:

- **BiLSTM:** Bidirectional Long Short-Term Memory (BiLSTM) (Schuster & Paliwal, 1997) combines two LSTM layers to capture textual information from both preceding and succeeding tokens.
- **BERT:** Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is a transformer-based language model that employs an encoder-only architecture to generate contextualized representations of input tokens. We utilize the BERT-BASE model in our experiments.



doi:10.6342/NTU202501600

- **FinBERT:** FinBERT (Huang et al., 2023; Yang et al., 2020) is a domain-specific adaptation of BERT, pre-trained on financial texts such as 10-K and 10-Q filings and analyst reports. This specialization allows it to better capture financial terminology and context.
- **SEC-BERT** and **SEC-BERT-NUM:** SEC-BERT (Loukas et al., 2022) is a BERT-based model pre-trained on SEC 10-K filings. We also adopt the SEC-BERT-NUM variant, which applies an additional masking strategy during pre-training. Specifically, it replaces all numerical values with a special token [NUM], allowing the model to treat numeric expressions more uniformly.

### 3.3.2 Task-specific Layer

Each task-specific output head includes a linear transformation followed by layer normalization, a GELU activation function, and dropout (with dropout rate depending on the attribute. A final linear layer maps the hidden representation to the output logits.

### 3.3.3 Loss Function

We applied a weighted loss function to balance the contributions of different attribute prediction tasks. The total loss is computed as:

$$\text{Loss} = \frac{1}{\Sigma w_{attr}} \Sigma (w_{attr} * Loss_{attr})$$

$$\text{attr} \in \{\text{tag,time,scale,negative}\}$$

where $Loss_{attr}$ is loss of specific attribute and $w_{attr}$ is weight of specific attribute. Based on the characteristics of each attribute, we apply different loss functions accordingly. The specific loss function used for each attribute is described as follows:

- Tag: Given the large number of tag classes and the severe class imbalance, the tag attribute exhibits a long-tail distribution. This long-tail distribution poses



doi:10.6342/NTU202501600

challenges for classification models, as they tend to favor frequent classes while underperforming on rare ones. Though we adopt **Class-Balanced Loss (CB Loss)** (Cui et al., 2019). CB Loss addresses data imbalance by re-weighting the loss function based on the *effective number of samples*, which considers the diminishing marginal benefit of additional samples in frequently occurring classes. The effective number of class y is defined as:

$$\mathrm{E}_y = \frac{1 - \beta^{\mathrm{n_y}}}{1 - \beta}$$

where $\mathrm{n_y}$ is the number of samples belonging to ground-truth class y, and $\beta \in [0,1)$ is a hyperparameter controlling the degree of balance. A higher $\beta$ (e.g., 0.99) places more emphasis on rare classes.

The CB loss reweights the standard cross-entropy losses using the inverse of the effective number of samples, thus tag loss function is resulted in:

$$\mathrm{Loss_{tag}} = \frac{1}{N}\sum_{i=1}^{N} \alpha_{y_i} * \mathcal{L}(\mathbf{p_i}, \mathrm{y_i})$$

$$\alpha_{y_i} = \frac{1 - \beta}{1 - \beta^{\mathrm{n}_{y_i}}}$$

where $N$ is the number of samples, $\mathbf{p_i}$ is predicted probability vector of sample $i$, $y_i$ is ground-truth label, and $\mathcal{L}(\mathbf{p_i}, y_i)$ is the standard cross-entropy loss.

- Time and Scale: Both time and scale attributes exhibit class imbalance. We apply standard **cross-entropy loss** with class weights inversely proportional to class frequency to mitigate this imbalance.
- Fact: The fact attribute represents numerical values with a wide range that includes negative numbers. Direct regression methods yield suboptimal performance due to this variability. Instead, we leverage the compositional nature of the fact attribute,



doi:10.6342/NTU202501600

which can be expressed as:

$$\text{Fact} = \text{numeric value of target text} \times 10^{\text{scale}} \times (-1)^{negative}$$

Accordingly, we extract the numerical component from the target text and introduce a separate binary classification task to predict the **negative** sign attribute.

- Negative: The negative attribute, which indicates whether a value is negative, is highly imbalanced, with approximately 97% of the training samples labeled as positive. This aligns with the design of XBRL taxonomies, which predominantly represent positive financial values. To address this issue, we employ **Focal Loss** (Lin et al., 2020), which has demonstrated effectiveness in handling class imbalance in various domain, such as computer vision and NLP datasets (Mukhoti et al., 2020). Focal Loss modifies the standard cross-entropy by introducing a modulating factor:

$$\text{Loss}_{\text{negative}} = -\alpha_t(1 - p_t)^{\gamma}\log(p_t)$$

where $p_t$ is the predicted probability of the true class, $\alpha_t$ is a weighting factor to balance positive and negative classes, and $\gamma$ is a focusing parameter that controls how much attention is paid to hard examples. When $\gamma = 0$, focal loss reduces to standard cross-entropy loss.



doi:10.6342/NTU202501600

# Chapter 4 Experiment

## 4.1 Dataset Summary Statistics

In order to provide a comprehensive understanding of our dataset, we conducted a series of summary statistical analyses across various dimensions, including dataset composition over time and subsets, attribute distributions, and textual context characteristics.

### 4.1.1 Overall Summary Statistics

Table 4 and Table 5 present the overall size and distribution of the dataset by year and across the training, validation, and test subsets. In total, the dataset comprises 113,303 filings, covering 6,665,760 instances and 12,125,232 annotated targets.

Table 6 summarizes the number of targets that include each type of attribute, reflecting that not all attributes are present for every instance. Table 7 provides the number of unique values for each attribute.

Table 4 Summary Statistics by Year

| Year | # of 10-K | # of 10-Q | # of Instances | # of Targets |
|---|---|---|---|---|
| 2019 | 2,322 | 5,552 | 578,122 | 1,064,277 |
| 2020 | 3,533 | 10,015 | 967,302 | 1,766,462 |
| 2021 | 6,209 | 17,140 | 1,304,767 | 2,369,010 |
| 2022 | 6,431 | 20,149 | 1,381,029 | 2,504,806 |
| 2023 | 5,959 | 18,806 | 1,415,253 | 2,571,070 |
| 2024 | 850 | 16,337 | 1,019,287 | 1,849,607 |
| Total | 25,304 | 87,999 | 6,665,760 | 12,125,232 |



doi:10.6342/NTU202501600

Table 5 Summary Statistics by Subset

| | Train | | Validation | | Test | | All |
|---|---|---|---|---|---|---|---|
| | # of | % | # of | % | # of | % | # of |
| # of Instances | 5,280,499 | 79.22% | 621,645 | 9.33% | 763,616 | 11.46% | 6,665,760 |
| # of Targets | 9,615,922 | 79.31% | 1,133,989 | 9.35% | 1,375,321 | 11.34% | 12,125,232 |
| Avg targets/ Instance | 1.8210 | | 1.8242 | | 1.8011 | | 1.8190 |

Table 6 Counts of Targets Contain Attribute

| | Train | | Validation | | Test | | All | |
|---|---|---|---|---|---|---|---|---|
| | # of | % | # of | % | # of | % | # of | % |
| Tag | 9,615,922 | 100.0% | 1,133,989 | 100.0% | 1,375,321 | 100.0% | 12,125,232 | 100.0% |
| Fact | 9,515,058 | 98.95% | 1,122,153 | 98.96% | 1,360,551 | 98.93% | 11,997,762 | 98.95% |
| Time | 9,593,603 | 99.77% | 1,131,384 | 99.77% | 1,371,067 | 99.69% | 12,096,054 | 99.76% |
| Measure | 9,615,922 | 100.0% | 1,133,989 | 100.0% | 1,375,321 | 100.0% | 12,125,232 | 100.0% |
| Decimals | 9,615,922 | 100.0% | 1,133,989 | 100.0% | 1,375,321 | 100.0% | 12,125,232 | 100.0% |
| Scale | 8,850,466 | 92.04% | 1,045,140 | 92.16% | 1,257,096 | 91.40% | 11,152,702 | 91.98% |
| Axis | 6,781,635 | 70.53% | 802,937 | 70.81% | 954,045 | 69.37% | 8,538,617 | 70.42% |
| Member | 6,781,635 | 70.53% | 802,937 | 70.81% | 954,045 | 69.37% | 8,538,617 | 70.42% |
| All targets | 9,615,922 | 100.0% | 1,133,989 | 100.0% | 1,375,321 | 100.0% | 12,125,232 | 100.0% |

Table 7 Unique Count of Attributes

| Attribute | Train | Validation | Test | All |
|---|---|---|---|---|
| Unique Tag | 420,157 | 158,949 | 177,711 | 446,338 |
| Unique Measure | 41,170 | 19,423 | 21,174 | 42,942 |
| Unique Decimals | 30 | 28 | 29 | 30 |
| Unique Scale | 20 | 17 | 17 | 20 |
| Unique Axis | 8,199 | 4,231 | 4,545 | 8,474 |
| Unique Member | 384,861 | 181,013 | 197,317 | 402,308 |



doi:10.6342/NTU202501600

### 4.1.2 Attribute Summary Statistics

#### 4.1.2.1 Tag

While the total number of unique tags in the dataset is 446,338, as shown in Table 8 and Table 9, only 7,579 are standard tags, representing approximately 1.70% of all unique tags. Despite this, standard tags account for about 72% of all target occurrences, indicating their dominance in practical usage.

Table 10 presents the most frequently occurring tags in the training, validation, and test subsets. To further address the long-tailed distribution of standard tags, we divided them into frequent and rare categories based on their occurrence in the training set. As shown in Table 11, 13% of the unique standard tags are classified as frequent (appearing more than 1,000 times), yet cover 86% of the total standard tag occurrences.

Table 8 Unique of Standard and Custom Tags

| Unique Tag | Train | | Validation | | Test | | All |
|---|---|---|---|---|---|---|---|
| | # of | % | # of | % | # of | % | # of |
| Standard Tag | 7,494 | 1.78% | 6,293 | 3.96% | 6,434 | 3.62% | 7,579 |
| Custom Tag | 412,663 | 98.22% | 152,656 | 96.04% | 171,277 | 96.38% | 438,759 |
| Total Unique | 420,157 | 100.0% | 158,949 | 100.0% | 177,711 | 100.0% | 446,338 |

Table 9 Counts of Standard and Custom Tags

| # of Tag | Train | Validation | Test | All | |
|---|---|---|---|---|---|
| | | | | # of | % |
| Standard Tag | 72.42% | 72.38% | 72.88% | 8,786,674 | 72.47% |
| Custom Tag | 27.58% | 27.62% | 27.12% | 3,338,558 | 27.53% |
| Total Count | 9,615,922 | 1,133,989 | 1,375,321 | 12,125,232 | 100.00% |



doi:10.6342/NTU202501600

Table 10 Top-10 Frequent Tag Distribution

| Tag Name | Train | Validation | Test |
|---|---|---|---|
| us-gaap:DebtInstrumentInterestRateStatedPercentage | 1.84% | 1.88% | 1.80% |
| us-gaap:DebtInstrumentBasisSpreadOnVariableRate1 | 1.48% | 1.48% | 1.44% |
| us-gaap:DebtInstrumentFaceAmount | 1.43% | 1.46% | 1.40% |
| us-gaap:LineOfCreditFacilityMaximumBorrowingCapacity | 1.38% | 1.39% | 1.37% |
| us-gaap:AllocatedShareBasedCompensationExpense | 1.12% | 1.11% | 1.12% |
| us-gaap:ConcentrationRiskPercentage1 | 0.93% | 0.94% | 0.94% |
| dei:EntityCommonStockSharesOutstanding | 0.76% | 0.72% | 1.67% |
| us-gaap:AmortizationOfIntangibleAssets | 0.71% | 0.71% | 0.71% |
| us-gaap:AntidilutiveSecuritiesExcludedFromComputationOfEarningsPerShareAmount | 0.70% | 0.68% | 0.70% |
| us-gaap:EffectiveIncomeTaxRateContinuingOperations | 0.64% | 0.62% | 0.65% |

Table 11 Unique and Counts of Frequent and Rare Tags in Training Set

| Standard Tag in Training Set | # of | Percentage |
|---|---|---|
| Unique Frequent Tag (Count ≥ 1,000) | 976 | 13.02 % |
| Unique Rare Tag (Count < 1,000) | 6,518 | 86.98 % |
| Total Unique Standard Tag | 7,494 | 100.0% |
| # of Frequent Tag (Count ≥ 1,000) | 6,001,283 | 86.18 % |
| # of Rare Tag (Count < 1,000) | 962,326 | 13.82 % |
| # of Total Standard Tag | 6,963,609 | 100.0% |

### 4.1.2.2 Fact

Table 12 provides a summary of the fact values in the dataset. These values span a wide range and include both negative and positive numbers, reflecting the diversity of financial quantities reported. To better illustrate the distribution, Figure 10 presents the log-transformed fact values, while Table 13 further breaks down the proportion of



doi:10.6342/NTU202501600

negative versus non-negative fact values. Table 14 and Figure 11 demonstrated the numeric value of target text distribution.

Table 12 Summary Statistics of Fact

| | Train | Validation | Test |
|---|---|---|---|
| mean | 4.39e + 12 | 2.70e + 12 | 7.51e + 11 |
| std | 2.19e + 15 | 1.64e + 15 | 8.57e + 14 |
| min | -1.17e + 12 | -3.46e + 11 | -8.27e + 11 |
| 25% | 5.00 | 5.00 | 5.20 |
| 50% | 1.33e + 6 | 1.35e + 6 | 1.40e + 6 |
| 75% | 2.50e + 7 | 2.55e + 7 | 2.55e + 7 |
| max | 2.56e + 18 | 1.00e + 18 | 1.00e + 18 |

Table 13 Distribution of Negative and Non-negative Fact Value

| Fact | Train | Validation | Test | All | |
|---|---|---|---|---|---|
| | | | | # of | % |
| Negative | 2.75% | 2.75% | 2.78% | 330,261 | 2.75% |
| Non-negative | 97.25% | 97.25% | 97.22% | 11,667,501 | 97.25% |
| Total Count | 9,515,058 | 1,122,153 | 1,360,551 | 11,997,762 | 100.0% |

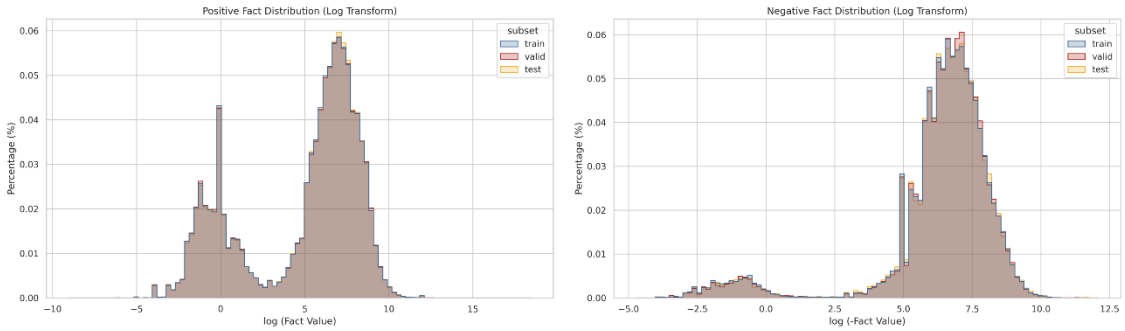



Figure 10 Positive and Negative Fact Value Distribution (Log Transform)



doi:10.6342/NTU202501600

Table 14 Summary Statistics of Numeric Value of Target Text

| | Train | Validation | Test |
|---|---|---|---|
| mean | 1.87e + 7 | 1.55e + 7 | 3.3e + 7 |
| std | 3.37e + 9 | 1.88e + 9 | 7.13e + 9 |
| min | 0.00 | 0.00 | 0.00 |
| 25% | 3.75 | 3.75 | 3.90 |
| 50% | 27.00 | 26.90 | 29.00 |
| 75% | 504.00 | 500.00 | 640.00 |
| max | 6e + 12 | 1.07e + 12 | 7.24e + 12 |

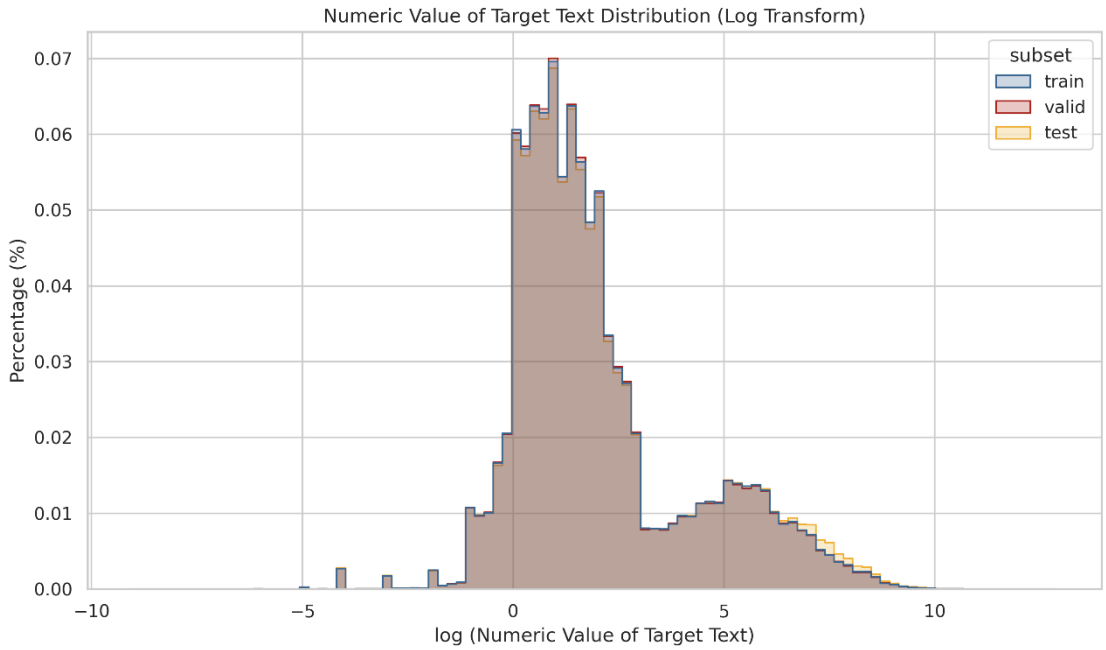



Figure 11 Numeric Value of Target Text Distribution (Log Transform)

### 4.1.2.3 Time

Table 15 presents the distribution of time attributes after being categorized into simplified classes. Most time references fall under either past or current, regardless of whether they represent an instant or a period.



doi:10.6342/NTU202501600

Table 15 Distribution of Time Category

| Time | Train | Validation | Test | All | |
|---|---|---|---|---|---|
| | | | | # of | % |
| instant; past | 18.67% | 18.72% | 18.43% | 2,255,336 | 18.65% |
| instant; current | 26.81% | 26.85% | 26.84% | 3,243,964 | 26.82% |
| instant; future | 1.97% | 1.93% | 2.78% | 248,873 | 2.06% |
| Total Instant | 47.45% | 47.51% | 48.05% | 5,748,173 | 47.52% |
| period; past | 24.09% | 24.01% | 23.68% | 2,907,069 | 24.03% |
| period; current | 17.22% | 17.31% | 16.87% | 2,079,432 | 17.19% |
| period; future | 1.48% | 1.51% | 1.46% | 178,945 | 1.48% |
| period; past_current | 9.65% | 9.55% | 9.81% | 1,167,837 | 9.65% |
| period; past_future | 0.07% | 0.07% | 0.08% | 8,729 | 0.07% |
| period; current_future | 0.05% | 0.06% | 0.05% | 5,869 | 0.05% |
| Total Period | 52.55% | 52.49% | 51.95% | 6,347,881 | 52.48% |
| Total Time Count | 9,593,603 | 1,131,384 | 1,371,067 | 12,096,054 | 100.00% |

### 4.1.2.4 Measure

Table 16 and Table 17 presents the distribution of the measure attribute, categorized into standard, custom, and divide types. Fraction-type measures consist of a numerator and denominator, each of which may use either a standard or custom unit. Notably, although standard measures constitute less than 1% of all unique categories, they account for over 90% of total occurrences in the dataset.

Table 18 further present that the top four most frequent measures alone contribute approximately 93% of all occurrences. This indicates a highly imbalanced distribution, with a small subset of categories dominating the measure attribute.



doi:10.6342/NTU202501600

Table 16 Unique of Standard, Custom and Fraction Measures

| Unique Measure | Train | | Validation | | Test | | All |
|---|---|---|---|---|---|---|---|
| | # of | % | # of | % | # of | % | # of |
| Standard Measure | 202 | 0.49% | 151 | 0.78% | 156 | 0.74% | 206 |
| Custom Measure | 40,537 | 98.46% | 19,056 | 98.11% | 20,775 | 98.12% | 42,293 |
| Fraction Measure | 431 | 1.05% | 216 | 1.11% | 243 | 1.15% | 443 |
| Total Unique | 41,170 | 100.0% | 19,423 | 100.0% | 21,174 | 100.0% | 42,942 |

Table 17 Count of Standard, Custom and Fraction Measures

| # of Measure | Train | Validation | Test | All | |
|---|---|---|---|---|---|
| | | | | # of | % |
| Standard Measure | 90.13% | 90.11% | 90.17% | 10,928,443 | 90.13% |
| Custom Measure | 5.07% | 5.07% | 5.11% | 615,060 | 5.07% |
| Fraction Measure | 4.80% | 4.83% | 4.72% | 581,729 | 4.80% |
| Total Count | 9,615,922 | 1,133,989 | 1,375,321 | 12,125,232 | 100.0% |

Table 18 Distribution of Measure

| Measure | Train | Validation | Test | All | |
|---|---|---|---|---|---|
| | | | | # of | % |
| iso4217:USD | 60.40% | 60.37% | 59.92% | 7,316,196 | 60.34% |
| xbrli:pure | 17.47% | 17.51% | 17.25% | 2,115,404 | 17.45% |
| xbrli:shares | 10.73% | 10.70% | 11.51% | 1,311,175 | 10.81% |
| iso4217:USD/xbrli:shares | 4.64% | 4.67% | 4.57% | 562,480 | 4.64% |
| other | 6.77% | 6.76% | 6.74% | 819,977 | 6.76% |
| Total Count | 9,615,922 | 1,133,989 | 1,375,321 | 12,125,232 | 100.0% |

### 4.1.2.5 Scale

The dataset contains only 20 unique scale values. Table 19 illustrates that scale value is 6, which corresponds to values in millions, accounts for nearly half of the dataset. This is followed by scale value is 0 (no scaling required) and -2 (values on the



doi:10.6342/NTU202501600

order of 0.01), which are also commonly used.

Table 19 Distribution of Scale

| Scale | Train | Validation | Test | All | |
|---|---|---|---|---|---|
| | | | | # of | % |
| 6 | 48.95% | 49.03% | 48.55% | 5,454,979 | 48.91% |
| 0 | 22.85% | 22.80% | 23.41% | 2,554,701 | 22.91% |
| -2 | 17.56% | 17.61% | 17.45% | 1,957,883 | 17.56% |
| 3 | 7.55% | 7.47% | 7.57% | 841,350 | 8.00% |
| 9 | 2.81% | 2.82% | 2.74% | 313,017 | 2.81% |
| -4 | 0.25% | 0.24% | 0.25% | 27,683 | 0.25% |
| other | 0.03% | 0.03% | 0.03% | 3,089 | 0.03% |
| Total | 8,850,466 | 1,045,140 | 1,257,096 | 11,152,702 | 100.0% |

### 4.1.2.6 Decimals

There is a total of 30 unique decimals value in dataset. Table 20 shows the distribution of the decimals attribute. The most frequent value is -5, indicating precision at a hundred thousand. The next most frequent is INF, which denotes that the value should be interpreted as fully precise (i.e., no rounding). This is followed by -6, indicating precision at the million level.



doi:10.6342/NTU202501600

Table 20 Distribution of Decimals

| Decimals | Train | Validation | Test | All | |
|---|---|---|---|---|---|
| | | | | # of | % |
| -5 | 31.14% | 31.24% | 30.82% | 3,772,640 | 31.11% |
| INF | 23.19% | 23.24% | 23.87% | 2,821,594 | 23.27% |
| -6 | 11.53% | 11.59% | 11.17% | 1,393,488 | 11.49% |
| 0 | 9.47% | 9.31% | 9.73% | 1,150,465 | 9.49% |
| -3 | 8.35% | 8.28% | 8.32% | 1,011,213 | 8.34% |
| 2 | 7.33% | 7.34% | 7.21% | 886,865 | 7.31% |
| 3 | 2.86% | 2.84% | 2.84% | 346,525 | 2.86% |
| 4 | 2.68% | 2.70% | 2.66% | 324,482 | 2.68% |
| -8 | 1.69% | 1.70% | 1.64% | 204,745 | 1.69% |
| other | 1.76% | 1.76% | 1.73% | 213,215 | 1.76% |
| Total | 9,615,922 | 1,133,989 | 1,375,321 | 12,125,232 | 100.0% |

### 4.1.2.7 Axis and Member

Table 21 illustrates the distribution of the number of axis or member attributes associated with each target (excluding targets with zero axis/member). Approximately 54% of targets are associated with a single axis or member, and 96% have fewer than three. However, in some rare cases, a target may be associated with up to eight axes or members. Table 22 to Table 25 report the total and unique counts of axis and member attributes, respectively. In the dataset, only 3.83% of axis attributes are standard, but they make up 97.78% of total occurrences. For member, greater proportion of unique members are custom: standard members account for 0.54% of unique values but contribute to 41% of the overall count.

Table 26 and Table 27 present the most frequently occurring axis and member attributes.



doi:10.6342/NTU202501600

Table 21 Number of Axis/Member per Target

| | Train | Validation | Test | All | |
|---|---|---|---|---|---|
| | | | | # of | % |
| 1 | 54.23% | 53.97% | 54.59% | 4,632,028 | 54.25% |
| 2 | 31.06% | 31.12% | 30.83% | 2,650,530 | 31.04% |
| 3 | 11.64% | 11.80% | 11.53% | 993,965 | 11.64% |
| 4 | 2.54% | 2.57% | 2.52% | 217,049 | 2.54% |
| 5 | 0.46% | 0.47% | 0.47% | 39,514 | 0.46% |
| 6 | 0.06% | 0.07% | 0.06% | 5,258 | 0.06% |
| 7 | 0.00% | 0.00% | 0.00% | 250 | 0.00% |
| 8 | 0.00% | 0.00% | 0.00% | 23 | 0.00% |
| Total | 6,781,635 | 802,937 | 954,045 | 8,538,617 | 100.0% |

Table 22 Unique of Standard and Custom Axis

| Unique Axis | Train | | Validation | | Test | | All |
|---|---|---|---|---|---|---|---|
| | # of | % | # of | % | # of | % | # of |
| Standard Axis | 314 | 3.83% | 305 | 7.21% | 304 | 6.69% | 314 |
| Custom Axis | 7,885 | 96.17% | 3,926 | 92.79% | 4,241 | 93.31% | 8,160 |
| Total Unique | 8,199 | 100% | 4,231 | 100% | 4,545 | 100% | 8,474 |

Table 23 Counts of Standard and Custom Axis

| # of Axis | Train | Validation | Test | All | |
|---|---|---|---|---|---|
| | | | | # of | % |
| Standard Axis | 97.78% | 97.75% | 97.85% | 13,703,940 | 97.79% |
| Custom Axis | 2.22% | 2.25% | 2.15% | 310,291 | 2.21% |
| Total Count | 11,131,054 | 1,322,159 | 1,561,018 | 14,014,231 | 100.0% |



doi:10.6342/NTU202501600

Table 24 Unique of Standard and Custom Member

| Unique Member | Train | | Validation | | Test | | All |
|---|---|---|---|---|---|---|---|
| | # of | % | # of | % | # of | % | # of |
| Standard Member | 2,063 | 0.54% | 1,724 | 0.95% | 1,757 | 0.89% | 2,090 |
| Custom Member | 382,798 | 99.46% | 179,289 | 99.05% | 195,560 | 99.11% | 400,218 |
| Total Unique | 384,861 | 100.0% | 181,013 | 100.0% | 197,317 | 100.0% | 402,308 |

Table 25 Counts of Standard and Custom Member

| # of Member | Train | Validation | Test | All | |
|---|---|---|---|---|---|
| | | | | # of | % |
| Standard Member | 41.65% | 41.51% | 41.99% | 5,840,197 | 41.67% |
| Custom Member | 58.35% | 58.49% | 58.01% | 8,174,034 | 58.33% |
| Total Count Member | 11,131,054 | 1,322,159 | 1,561,018 | 14,014,231 | 100.0% |

Table 26 Top-10 Frequent Axis Distribution

| Axis Name | Train | Validation | Test |
|---|---|---|---|
| us-gaap:DebtInstrumentAxis | 9.37% | 9.41% | 9.32% |
| us-gaap:LongtermDebtTypeAxis | 6.73% | 6.78% | 6.76% |
| dei:LegalEntityAxis | 5.33% | 5.40% | 5.13% |
| us-gaap:AwardTypeAxis | 4.58% | 4.53% | 4.67% |
| us-gaap:BusinessAcquisitionAxis | 3.80% | 3.84% | 3.78% |
| us-gaap:CreditFacilityAxis | 3.67% | 3.68% | 3.69% |
| us-gaap:RelatedPartyTransactionsByRelatedPartyAxis | 3.63% | 3.62% | 3.61% |
| us-gaap:StatementClassOfStockAxis | 3.62% | 3.59% | 3.80% |
| srt:RangeAxis | 3.33% | 3.32% | 3.30% |
| us-gaap:PlanNameAxis | 2.54% | 2.54% | 2.56% |



doi:10.6342/NTU202501600

Table 27 Top-10 Frequent Member Distribution

| Member Name | Train | Validation | Test |
|---|---|---|---|
| us-gaap:RevolvingCreditFacilityMember | 2.08% | 2.11% | 2.11% |
| us-gaap:SubsequentEventMember | 1.93% | 1.93% | 1.93% |
| srt:MaximumMember | 1.85% | 1.85% | 1.84% |
| us-gaap:LineOfCreditMember | 1.55% | 1.55% | 1.58% |
| srt:MinimumMember | 1.43% | 1.42% | 1.41% |
| us-gaap:SeniorNotesMember | 1.24% | 1.25% | 1.26% |
| us-gaap:CommonClassAMember | 1.00% | 0.99% | 1.17% |
| us-gaap:RestrictedStockUnitsRSUMember | 0.96% | 0.94% | 0.98% |
| us-gaap:CommonStockMember | 0.89% | 0.91% | 0.89% |
| us-gaap:SecuredDebtMember | 0.75% | 0.76% | 0.75% |

### 4.1.3 Textual Context Summary Statistics

Figure 12 and Table 28 summarize the length distribution and descriptive statistics of the textual context across different subsets, presented at both the character level and token level. The tokenization was performed using the BERT-BASE tokenizer. We also report the proportion of context segments that exceed or fall below 512 tokens, which is the maximum sequence length supported by BERT-based models.



doi:10.6342/NTU202501600

Table 28 Summary Statistics of Context Length

| Context Type | Metrics | Train | Validation | Test |
|---|---|---|---|---|
| context_p (character level) | Average | 191.6787 | 191.8631 | 191.2571 |
| | Min | 8 | 11 | 11 |
| | Max | 5,735 | 5,293 | 3,527 |
| context_t (character level) | Average | 212.9535 | 213.2620 | 211.5472 |
| | Min | 10 | 10 | 10 |
| | Max | 8,689 | 6,866 | 6,837 |
| context_n (character level) | Average | 172.9194 | 173.0169 | 172.5690 |
| | Min | 8 | 9 | 9 |
| | Max | 6,832 | 3,622 | 3,699 |
| context_p (token level) | Average | 40.86 | 40.89 | 40.77 |
| | Min | 3 | 4 | 4 |
| | Max | 1,304 | 1,186 | 765 |
| | > 512 tokens | 292 (0.01%) | 33 (0.01%) | 34 (0.01%) |
| | ≤ 512 tokens | 3,240,117 (99.99%) | 381,754 (99.99%) | 457,948 (99.99%) |
| context_t (token level) | Average | 47.75 | 47.81 | 47.60 |
| | Min | 4 | 4 | 4 |
| | Max | 1,788 | 1,456 | 1,743 |
| | > 512 tokens | 942 (0.02%) | 120 (0.02%) | 150 (0.02%) |
| | ≤ 512 tokens | 5,279,557 (99.98%) | 621,525 (99.98%) | 763,466 (99.98%) |
| context_n (token level) | Average | 36.59 | 36.60 | 36.50 |
| | Min | 3 | 4 | 4 |
| | Max | 1,449 | 1,000 | 761 |
| | > 512 tokens | 104 (0.00%) | 9 (0.00%) | 10 (0.00%) |
| | ≤ 512 tokens | 3,375,368 (100.00%) | 397,590 (100.00%) | 478,640 (100.00%) |



doi:10.6342/NTU202501600

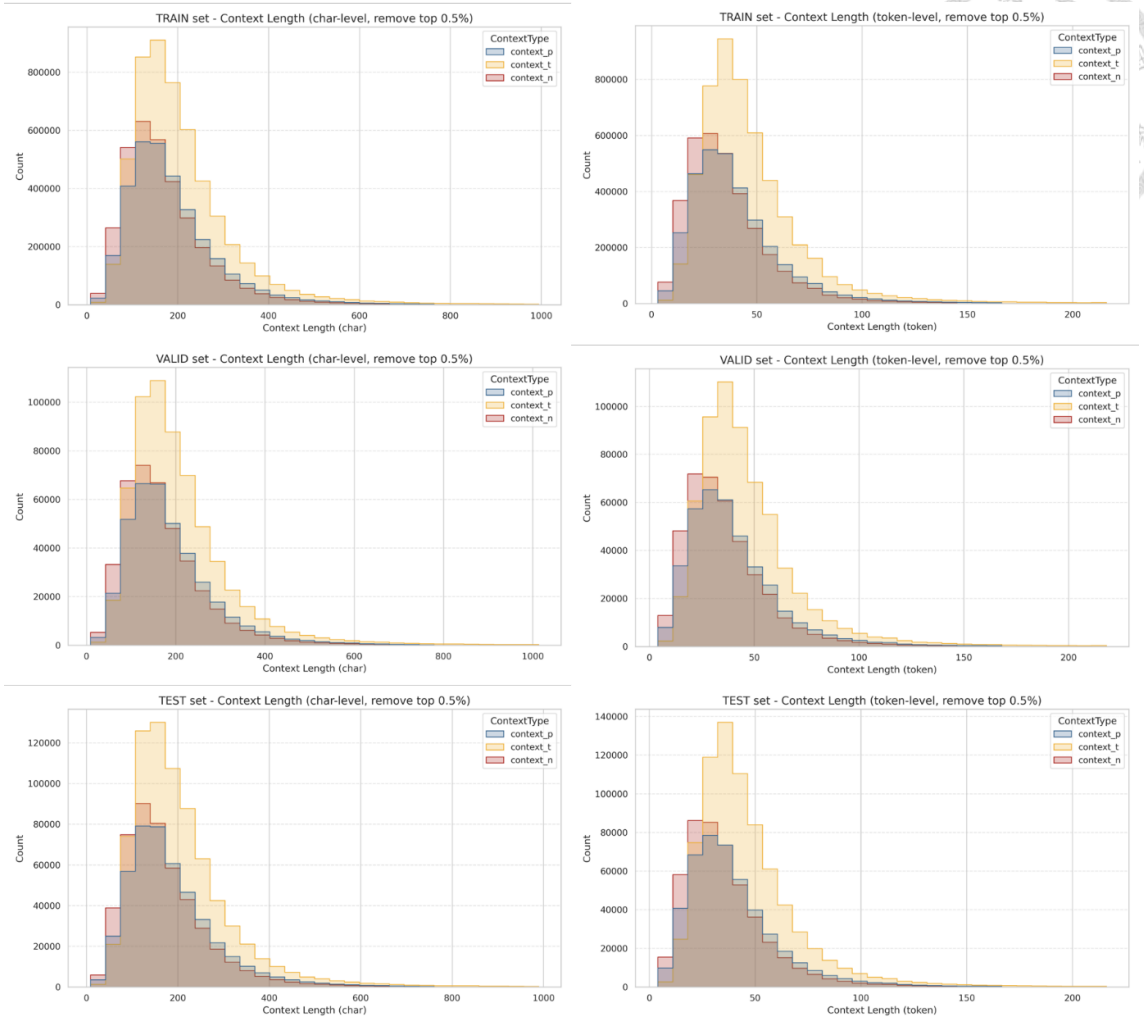



Figure 12 Context Length Distribution (Character Level and Token Level)

## 4.2 Experiment Implementation

In this section, we provide details of the experiment implemented in both multi-task learning models and few-shot learning using LLMs.

### 4.2.1 Multi-task Learning Model

#### 4.2.1.1 Data Preparation

Model training is conducted at the target level. For each instance, the target sequence is first extracted, and the input sequence is constructed by concatenating



doi:10.6342/NTU202501600

context_t, context_p, context_n, and document_info, separated by special tokens.

For BERT-based model, tokenization is performed using the tokenizer associated with each backbone model. Among them, for SEC-BERT-NUM, all numerical values in the context are replaced with the special token [NUM] prior to tokenization. For the BiLSTM model, tokenization is performed using spaCy. The vocabulary is filtered to include tokens with frequency greater than 2, and all numeric expressions are replaced with <NUM>, resulting in a final vocabulary size of 97,918.

#### 4.2.1.2 Model Architecture

For the BiLSTM model, we use a two-layer BiLSTM with an embedding dimension of 256 and a hidden size of 768. The final representation is obtained by concatenating the last forward and backward hidden states. Dropout is applied with a rate of 0.3 for the BiLSTM encoder, as well as the tag and scale heads, and 0.5 for time and negative heads.

For BERT-based models, the final hidden states of the backbone are extracted. The embedding corresponding to the position of the target span is used as the target embedding, which is then passed to the attribute-specific prediction heads. A dropout rate of 0.3 is applied to the tag and time heads, and 0.5 to scale and negative head.

The AdamW optimizer and is CosineAnnealingLR schedular used across all models. Additional hyperparameter details are summarized in Table 29.



doi:10.6342/NTU202501600

Table 29 Hyper-parameter Setting

| Backbone | Head | Learning Rate | Weight Decay | Weight | Batch size |
|---|---|---|---|---|---|
| BiLSTM | Backbone | 1e-3 | 1e-5 | - | 512 |
| | Tag | 5e-4 | 5e-5 | 10.0 | |
| | Time | 3e-4 | 5e-5 | 0.3 | |
| | Scale | 3e-4 | 5e-5 | 0.2 | |
| | Negative | 2e-4 | 5e-5 | 8.0 | |
| BERT | Backbone | 1e-5 | 1e-2 | - | 64 |
| | Tag | 5e-4 | | 10.0 | |
| | Time | 1e-5 | | 0.3 | |
| | Scale | 3e-5 | | 0.2 | |
| | Negative | 2e-5 | | 10.0 | |
| FinBERT | Backbone | 1e-4 | 1e-2 | - | 128 |
| | Tag | 1e-3 | | 10.0 | |
| | Time | 5e-5 | | 0.3 | |
| | Scale | 1e-4 | | 0.2 | |
| | Negative | 5e-6 | | 10.0 | |
| SEC-BERT | Backbone | 1e-5 | 1e-2 | - | 64 |
| | Tag | 5e-4 | | 10.0 | |
| | Time | 1e-5 | | 0.3 | |
| | Scale | 3e-5 | | 0.2 | |
| | Negative | 2e-5 | | 10.0 | |
| SEC-BERT-NUM | Backbone | 1e-5 | 1e-2 | - | 128 |
| | Tag | 5e-4 | | 10.0 | |
| | Time | 1e-5 | | 0.3 | |
| | Scale | 3e-5 | | 0.2 | |
| | Negative | 5e-5 | | 10.0 | |



doi:10.6342/NTU202501600

#### 4.2.1.3 Loss Functions

The detail of loss functions for different attributes are customized are following:

- Scale: Class weights are applied such that the classes with values 3 and 9 are assigned weights of 1.2 and 1.5, respectively, while the remaining classes retain a weight of 1.
- Time: Class weights are computed based on class frequency, followed by a logarithmic transformation. A minimum weight of 1 is enforced to prevent weights from becoming too small.
- Negative: Focal Loss is applied with parameters $\alpha = 0.25$ and $\gamma = 3.0$ to address the severe class imbalance in the negative class.
- Tag: Class-Balanced Cross Entropy Loss (CB-CE) is adopted with $\beta = 0.99$, considering the long-tailed distribution of tag classes.

A weighted sum of the loss functions from each task is used to optimize the multi-task learning model. These weights are manually adjusted based on each task's convergence behavior and the relative scale of individual losses. This strategy helps balance the learning progress across tasks and prevents certain tasks with disproportionately large losses from dominating the optimization process. It also helps to maintain comparable learning speeds among different attribute prediction tasks. The weighted setting is shown in Table 29.

### 4.2.2 Few-shots Learning using LLMs

We also conducted a series of few-shot experiments using LLMs, including GPT-4o, GPT-3.5-turbo, and Gemini 2.0 Flash. A total of 1,000 instances were randomly selected from the test set, and filtered to retain only those containing all four target attributes: tag, time, scale, and fact. Furthermore, to specifically evaluate the



doi:10.6342/NTU202501600

model's ability to generate tag names, we further filtered the instance labeled with frequent standard tags. The final evaluation set consisted of 976 samples, covering 352 unique standard tags. We augmented the prompt with four example input-output pairs to guide the model. The detail of prompt is provided in the Appendix.

### 4.2.3 Evaluation Metrics

The classification-based attributes are evaluated using both macro-averaged and weighted variants of common metrics: F1 score, precision, recall. Given the significant class imbalance observed in several attributes, weighted metrics are employed to provide a more balanced assessment of model performance across frequent and infrequent classes. Furthermore, due to the large number of possible tag labels, Hits@k is used to evaluate whether the correct tag appears within the top $k$ predictions, offering insight into the model's ranking capability.

To evaluate the ability of LLMs to generate tag names in a few-shot setting, we use BLEU scores (Papineni et al., 2002), which measure the similarity between generated outputs and reference tags. We also utilize Jaccard similarity, which captures the overlap between the sets of tokens in the predicted and reference tags, providing a complementary view of semantic similarity based on token-level agreement.

As mentioned before, the fact attribute presents additional challenges due to its large numerical variation and potential negative values. Directly applying regression metrices such as Mean Squared Error (MSE) may not be appropriate in this context. Therefore, we decompose the fact prediction into scale and negative sign prediction, which are evaluated using the same classification-based metrics described above.



doi:10.6342/NTU202501600

## 4.3 Experiment Result

### 4.3.1 Result of Multi-task Learning Model

The results of fine-tuned model are summarized in Table 30 and Table 31. In terms of weighted F1 scores, SEC-BERT achieved the highest overall performance across most attributes, with scores of 82.60 for tag prediction, 90.12 for time, 99.27 for scale, and 99.12 for negative sign prediction. The strong performance on scale and negative attributes reflects the model's ability to accurately predict fact attributes. Notably, FinBERT followed closely and even outperformed SEC-BERT on time attribute, achieving a weighted F1 score of 90.16. These results demonstrate the effectiveness of domain-specific pretraining in improving model performance on XBRL-related tasks. In contrast, the comparatively weaker performance of SEC-BERT-NUM, particularly on tag and time prediction, suggests that simply incorporating numerical representation does not guarantee improved outcomes. We also report Hits@k metrics for the tag prediction to evaluate the model's ability to rank relevant tags. As shown in Table 32, SEC-BERT outperforms others with a Hits@5 of 97.43%, indicating its strong ability to position the correct tag among the top predicted candidates.

Table 30 Result of Weighted Metrics

| Weighted | Tag | | | Time | | | Scale | | | Negative | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | R | P | F1 | R | P | F1 | R | P | F1 | R | P |
| BiLSTM | 71.66 | 72.19 | 71.82 | 64.97 | 65.30 | 67.93 | 91.09 | 91.11 | 91.16 | 97.84 | 98.19 | 98.06 |
| BERT | 78.95 | 79.47 | 79.29 | 89.43 | 89.41 | 89.54 | 99.23 | 99.24 | 99.23 | 98.99 | 99.04 | 98.99 |
| FinBERT | 81.50 | 81.80 | 81.67 | **90.16** | **90.16** | **90.21** | 99.25 | 99.26 | 99.25 | 99.01 | 99.06 | 99.01 |
| SEC-BERT | **82.60** | **82.81** | **82.73** | 90.12 | 90.12 | 90.15 | **99.27** | **99.27** | **99.26** | **99.12** | **99.13** | **99.12** |
| SEC-BERT-NUM | 76.18 | 76.83 | 76.77 | 83.08 | 83.01 | 83.38 | 98.40 | 98.40 | 98.40 | 98.91 | 98.96 | 98.91 |



doi:10.6342/NTU202501600

Table 31 Result of Macro Metrics

| Macro | Tag | | | Time | | | Scale | | | Negative | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | R | P | F1 | R | P | F1 | R | P | F1 | R | P |
| BiLSTM | 61.79 | 65.18 | 60.46 | 60.21 | 59.61 | 63.25 | 38.80 | 45.16 | 37.06 | 76.85 | 94.17 | 69.54 |
| BERT | 70.07 | 75.30 | 67.82 | 81.83 | 80.55 | 83.33 | 48.15 | 53.65 | 46.37 | 90.32 | 94.21 | 87.09 |
| FinBERT | 73.54 | 76.89 | 71.92 | **83.46** | **83.69** | **83.30** | 49.90 | **65.90** | 47.25 | 90.44 | **94.88** | 86.85 |
| SEC-BERT | **75.15** | **77.80** | **73.73** | 82.80 | 83.09 | 82.57 | **51.13** | 59.00 | **48.74** | **91.85** | 92.15 | **91.55** |
| SEC-BERT-NUM | 66.18 | 71.48 | 64.86 | 75.05 | 72.96 | 77.71 | 43.63 | 45.69 | 42.73 | 89.45 | 93.86 | 85.89 |

Table 32 Tag Hits@k

| Hits@k | hits@1 | hits@3 | hits@5 |
|---|---|---|---|
| BiLSTM | 72.18 | 92.11 | 95.97 |
| BERT | 79.46 | 94.02 | 97.11 |
| FinBERT | 81.79 | 94.62 | 97.32 |
| SEC-BERT | **82.81** | **94.95** | **97.43** |
| SEC-BERT-NUM | 76.83 | 93.20 | 96.72 |

### 4.3.2 Result of Few-shot Learning

The results of few-shot learning are presented in Table 33 and Table 34. For comparison, we also included results from our BERT and SEC-BERT model. Among the LLMs evaluated, the GPT-4o achieved the highest performance across the tag and negative attributes, with weighted F1 scores of 24.13 and 97.06, respectively. In contrast, Gemini 2.0 Flash outperformed other models on scale attribute, achieving a weighted F1 score of 87.93, though its performance on tag attribute was relatively weak. While LLMs demonstrated relatively acceptable results for scale and negative attribute, their overall ability to perform multi-attribute classification using few-shot prompting was significantly lower than that of a task-specific fine-tuned model.

The BLEU and n-gram precision scores on tag attribute similarly reflect this



doi:10.6342/NTU202501600

conclusion. SEC-BERT achieved a BLEU score of 87.50 with consistently high precision across all n-grams. In comparison, the best-performing LLM (GPT-4o, temperature = 0) reached a BLEU score of 46.60, while all other LLM variants scored below 24. These results suggest that, while LLMs possess general reasoning abilities, their performance on domain-specific classification tasks remains limited without fine-tuning.

Table 33 Result of Few-shot on LLMs

| Weighted | Tag | | | Time | | | Scale | | | Negative | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **F1** | **R** | **P** | **F1** | **R** | **P** | **F1** | **R** | **P** | **F1** | **R** | **P** |
| BERT | 78.73 | 78.48 | 81.49 | 91.80 | 91.80 | 91.94 | **99.39** | **99.39** | **99.39** | 99.02 | **99.08** | **99.09** |
| SEC-BERT | **81.81** | **80.94** | **85.46** | **92.96** | **92.93** | **93.13** | 99.38 | **99.39** | **99.39** | **99.03** | **99.08** | 99.06 |
| GPT-4o (temperature = 0) | 24.13 | 24.39 | 28.93 | 58.49 | 58.20 | 61.61 | 87.28 | 87.40 | 87.58 | 97.06 | 97.13 | 97.00 |
| GPT-4o (temperature = 0.5) | 23.46 | 23.16 | 31.60 | 58.26 | 57.27 | 62.10 | 87.47 | 87.40 | 87.88 | 97.06 | 97.13 | 97.00 |
| GPT-4o (temperature = 1) | 19.87 | 18.95 | 29.04 | 56.85 | 56.05 | 60.21 | 87.54 | 87.50 | 87.96 | 96.93 | 97.03 | 96.86 |
| GPT-4o mini (temperature = 0) | 7.49 | 9.32 | 8.29 | 36.76 | 38.11 | 44.42 | 73.17 | 73.46 | 76.16 | 94.89 | 94.36 | 95.57 |
| GPT-3.5 Turbo (temperature = 0) | 11.06 | 12.40 | 15.41 | 28.12 | 28.18 | 36.54 | 74.62 | 73.26 | 79.42 | 95.04 | 94.47 | 95.83 |
| Gemini 2.0 Flash (temperature = 0) | 9.37 | 9.94 | 11.91 | 58.95 | 59.12 | 61.00 | 87.09 | 86.17 | 88.28 | 96.39 | 96.41 | 96.37 |
| Gemini 2.0 Flash (temperature = 1) | 9.61 | 10.04 | 12.81 | 58.20 | 58.71 | 60.46 | 87.93 | 86.89 | 89.20 | 96.60 | 96.52 | 96.69 |
| Gemini 2.0 Flash-Lite (temperature = 0) | 9.44 | 10.04 | 11.97 | 59.28 | 59.32 | 61.43 | 87.03 | 86.07 | 88.23 | 96.31 | 96.31 | 96.31 |
| Gemini 2.0 Flash-Lite (temperature = 1) | 9.96 | 10.25 | 12.70 | 58.29 | 58.50 | 60.65 | 86.62 | 85.66 | 87.81 | 96.31 | 96.31 | 96.31 |



doi:10.6342/NTU202501600

Table 34 BLEU score and Jaccard Similarity

| Model | BLEU | Jaccard Similarity |
|---|---|---|
| BERT | 86.13 | 84.38 |
| SEC-BERT | **87.50** | **85.92** |
| GPT-4o (temperature = 0) | <u>46.60</u> | <u>55.78</u> |
| GPT-4o (temperature = 0.5) | 45.14 | 55.00 |
| GPT-4o (temperature = 1) | 41.53 | 51.43 |
| GPT-4o mini (temperature = 0) | 17.47 | 39.61 |
| GPT-3.5 Turbo (temperature = 0) | 23.67 | 40.72 |
| Gemini 2.0 Flash (temperature = 0) | 17.41 | 40.88 |
| Gemini 2.0 Flash (temperature = 1) | 17.59 | 40.58 |
| Gemini 2.0 Flash-Lite (temperature = 0) | 17.46 | 40.91 |
| Gemini 2.0 Flash-Lite (temperature = 1) | 16.55 | 39.67 |

## 4.4 Error Analysis

### 4.4.1 Tag

The error analysis is conducted based on the prediction results of the BERT backbone model. Among the tag predictions, Table 35 presents the most frequently misclassified classes and the class pairs that are most often confused. A notable source of misclassification is between custom and standard_rare tags, which are frequently confused with one another. In addition, some standard tags are frequently misclassified, particularly when their semantic meanings are closely related. For example, tags associated with debt instruments (LineOfCreditFacilityMaximumBorrowingCapacity, DebtInstrumentCarryingAmount, and DebtInstrumentFaceAmount) are often confused with each other. Similar issues occur with tags related to share prices (SharePrice and SharesIssuedPricePerShare) and tags related to stock issuances



doi:10.6342/NTU202501600

(SaleOfStockNumberOfSharesIssuedInTransaction and StockIssuedDuringPeriodSharesNewIssues). These misclassifications may not solely result from model errors, but could also be attributed to inconsistencies or inaccuracies in the original XBRL tagging, which introduce noise into the training data. Figure 13 demonstrated several error prediction examples on tag attribute.

Table 36 presents the tag classes with the highest false positive rates, highlighting the tags that are more likely to be over-predicted. Some tags tend to be misclassified as custom. This may be due to inconsistencies in reporting, where companies use custom tags in situations where standard concepts would be more appropriate.

Table 35 Most Frequent Tag Misclassified Class

| True Tag | Total Count | Total Error | Most Frequent Misclassified Tag | Error Pair Count | Pair / Total Error |
|---|---|---|---|---|---|
| custom | 372,750 | 56,673 | standard_rare | 14,435 | 25.47% |
| standard_rare | 136,529 | 40,006 | custom | 18,349 | 45.87% |
| us-gaap:DebtInstrumentFaceAmount | 19,208 | 2,997 | us-gaap:LineOfCreditFacilityMaximumBorrowingCapacity | 917 | 30.60% |
| us-gaap:StockIssuedDuringPeriodSharesNewIssues | 7,936 | 2,616 | custom | 1,114 | 42.58% |
| | | | us-gaap:SaleOfStockNumberOfSharesIssuedInTransaction | 618 | 23.62% |
| us-gaap:DebtInstrumentCarryingAmount | 4,876 | 2,525 | us-gaap:DebtInstrumentFaceAmount | 1,172 | 46.42% |
| us-gaap:SharePrice | 3,856 | 2,235 | custom | 811 | 36.29% |
| | | | us-gaap:SharesIssuedPricePerShare | 706 | 31.59% |
| us-gaap:SaleOfStockNumberOfSharesIssuedInTransaction | 4,045 | 2,138 | us-gaap:StockIssuedDuringPeriodSharesNewIssues | 1,305 | 61.04% |



doi:10.6342/NTU202501600

Table 36 Tag Class False Positive Rate

| True Tag | False Positive Rate |
|---|---|
| custom | 6.64% |
| standard_rare | 2.60% |
| us-gaap:DebtInstrumentFaceAmount | 0.49% |
| us-gaap:LineOfCreditFacilityMaximumBorrowingCapacity | 0.31% |
| us-gaap:AllocatedShareBasedCompensationExpense | 0.25% |
| us-gaap:StockIssuedDuringPeriodSharesNewIssues | 0.23% |

**Example 1**

| | |
|---|---|
| context_t | The facility allows the Company to borrow, subject to compliance with borrowing base requirements and other conditions, up to $10,000,000 to facilitate the acquisition of multi-family properties, and is secured by the cash available in certain cash accounts maintained by the Company at Valley National Bank. |
| true tag | **us-gaap:DebtInstrumentFaceAmount** |
| pred tag | **us-gaap:LineOfCreditFacilityMaximumBorrowingCapacity** |

**Example 2**

| | |
|---|---|
| context_t | In the first quarter of 2019, we issued $300 million of 5.375% Senior Notes due 2049 at a discount of $5.9 million which, when coupled with debt issuance costs of approximately $3.3 million, resulted in net proceeds from the offering of $290.8 million. |
| true tag | **us-gaap:DebtInstrumentCarryingAmount** |
| pred tag | **us-gaap:DebtInstrumentFaceAmount** |

**Example 3**

| | |
|---|---|
| context_t | The common stock was issued at $34.62 per share and thus the Private Placement noted above was priced at $34.62 per share. |
| true tag | **us-gaap:SharePrice** |
| pred tag | **us-gaap:DebtInstrumentFaceAmount** |

**Example 4**

| | |
|---|---|
| context_t | In June 2019, we sold 12.7 million shares of our common stock under a registered public offering for gross proceeds of $62.75 per share. |
| true tag | **us-gaap:StockIssuedDuringPeriodSharesNewIssues** |
| pred tag | **us-gaap:SaleOfStockNumberOfSharesIssuedInTransaction** |

Figure 13 Tag Error Example

### 4.4.2 Time

For time attribute, Figure 14 presents the confusion matrix of time attribute predicted result. Most classes achieved satisfactory performance; however, classed such as period:future, period:current_future and period:past_future posed greater



doi:10.6342/NTU202501600

challenge, which may due to their low frequency in dataset.

Table 37 and Table 38 further detail the most commonly misclassified time classes, the frequently confused class pairs, and the false positive rate for each class. It is worth noting that although misclassifications exist, the most frequent confused classes often share at least one dimension with the true label, either in the type (period or instant) or the temporal context (past, current, or future). This suggests that the model captures partial information correctly even in error cases.

Table 37 Time Misclassified Class

| True Time | Total Count | Total Error | Most Frequent Misclassified Time | Error Pair Count | Pair / Total Error |
|---|---|---|---|---|---|
| instant; current | 2,571,644 | 42,600 | instant; past | 25,719 | 60.37% |
| period; current | 1,651,906 | 31,088 | period; past | 13,240 | 42.59% |
| period; past | 2,310,283 | 28,280 | instant; past | 10,432 | 36.89% |
| instant; past | 1,790,496 | 20,064 | instant; current | 11,673 | 58.18% |
| period; past_current | 925,158 | 13,431 | period; past | 5,762 | 42.90% |
| instant; future | 188,805 | 4,780 | period; future | 1,823 | 38.14% |
| period; future | 141,793 | 4,434 | instant; future | 1,421 | 32.05% |
| period; current_future | 4,607 | 278 | period; future | 85 | 30.58% |
| period; past_future | 6,837 | 221 | period; past_current | 63 | 28.51% |

Table 38 Time Class False Positive Rate

| True Time | False Positive Rate |
|---|---|
| instant; past | 3.59% |
| period; past | 2.71% |
| instant; current | 2.70% |
| period; current | 2.00% |
| period; past_current | 1.48% |



doi:10.6342/NTU202501600

| True Time | False Positive Rate |
|---|---|
| period; future | 0.44% |
| instant; future | 0.41% |
| period; past_future | 0.03% |
| period; current_future | 0.02% |

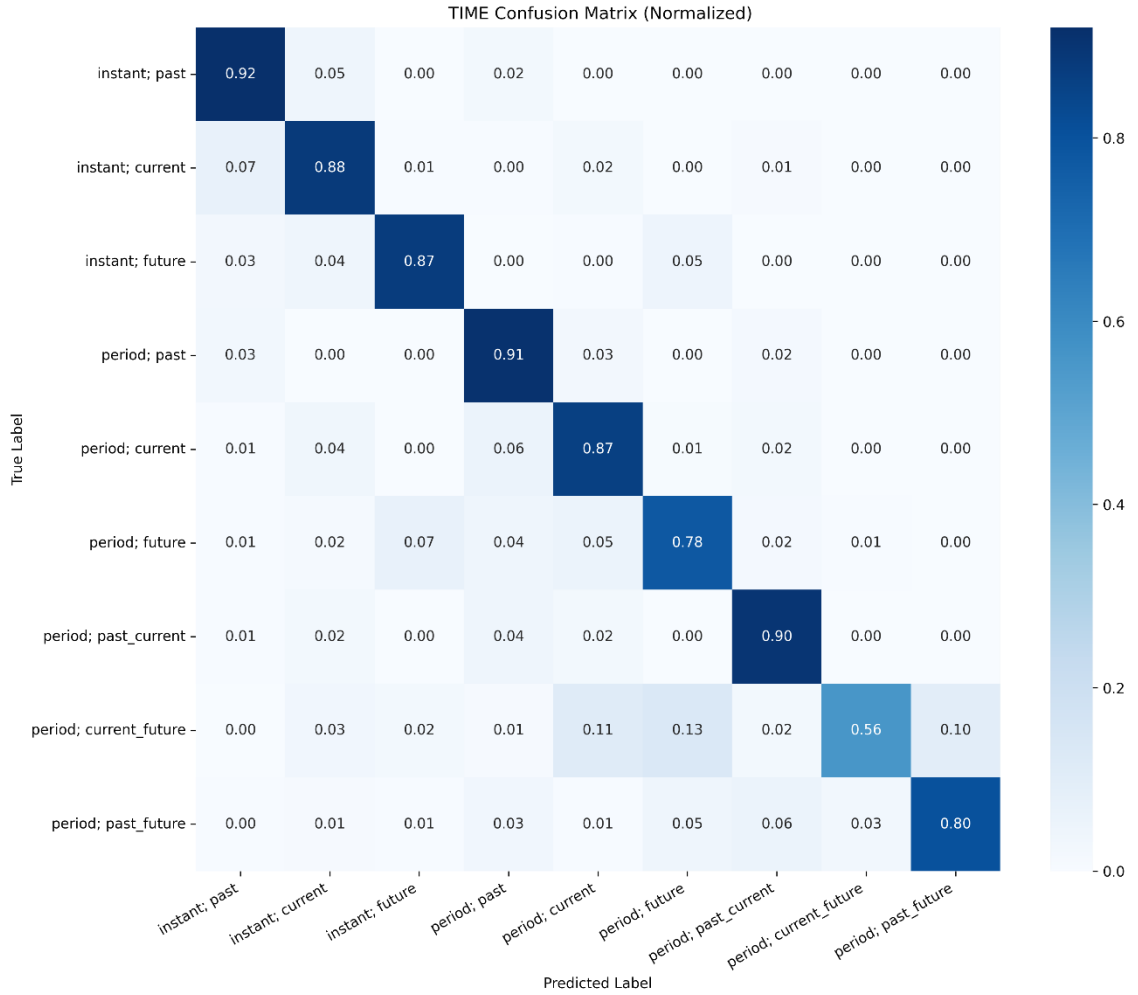



Figure 14 Confusion Matrix of Time Prediction Result (Normalized)

### 4.4.3 Scale

Additionally, Figure 15 presents the confusion matrix of the scale attribute predictions, while Table 39 and



doi:10.6342/NTU202501600

Table 40 provide detailed information on misclassified instances. Notably, several classes are frequently misclassified as class ‘0’. In contrast, certain classes (such as ‘6’ or ‘-2’) are more consistently predicted when specific textual patterns appear, such as the presence of spans like “millions” or symbols like “%”, which provide strong contextual clues.

In addition, there was a notable inconsistency between the macro and weighted scores as mentioned in Table 30 and Table 31, which can be attributed to several scale classes had fewer than 30 instances in the test set. Upon closer inspection, we identified two major causes of prediction errors. First, if the related scale context is too far from the target sentence, our dataset doesn't include this information. For example, the relevant information indicating the scale (e.g., "presented in thousands" or "in millions of shares") appear at the beginning of a note or in a distant part of the document, beyond the scope of the available context. Second, we discovered potential annotation errors in the dataset. In some cases, such as "5 million" being annotated with a scale of 7 (indicating a fact value of 50,000,000), our model predicted a scale of 6, which is more accurate. Figure 16 illustrate some example.

Table 39 Most Frequent Scale Misclassified Class

| True Scale | Total Count | Total Error | Most Frequent Misclassified Scale | Error Pair Count | Pair / Total Error |
|---|---|---|---|---|---|
| 0 | 294,091 | 4,047 | 3 | 1,979 | 48.90% |
| 3 | 95,106 | 2,911 | 0 | 2,123 | 72.93% |
| 6 | 610,255 | 1,154 | 0 | 608 | 52.69% |
| -2 | 219,297 | 907 | 0 | 555 | 61.19% |
| -4 | 3,086 | 181 | -2 | 166 | 91.71% |



doi:10.6342/NTU202501600

Table 40 Scale Class False Positive Rate

| True Scale | False Positive Rate |
|---|---|
| 0 | 10.80% |
| 3 | 0.46% |
| 6 | 0.29% |
| -2 | 0.20% |
| -4 | 0.02% |

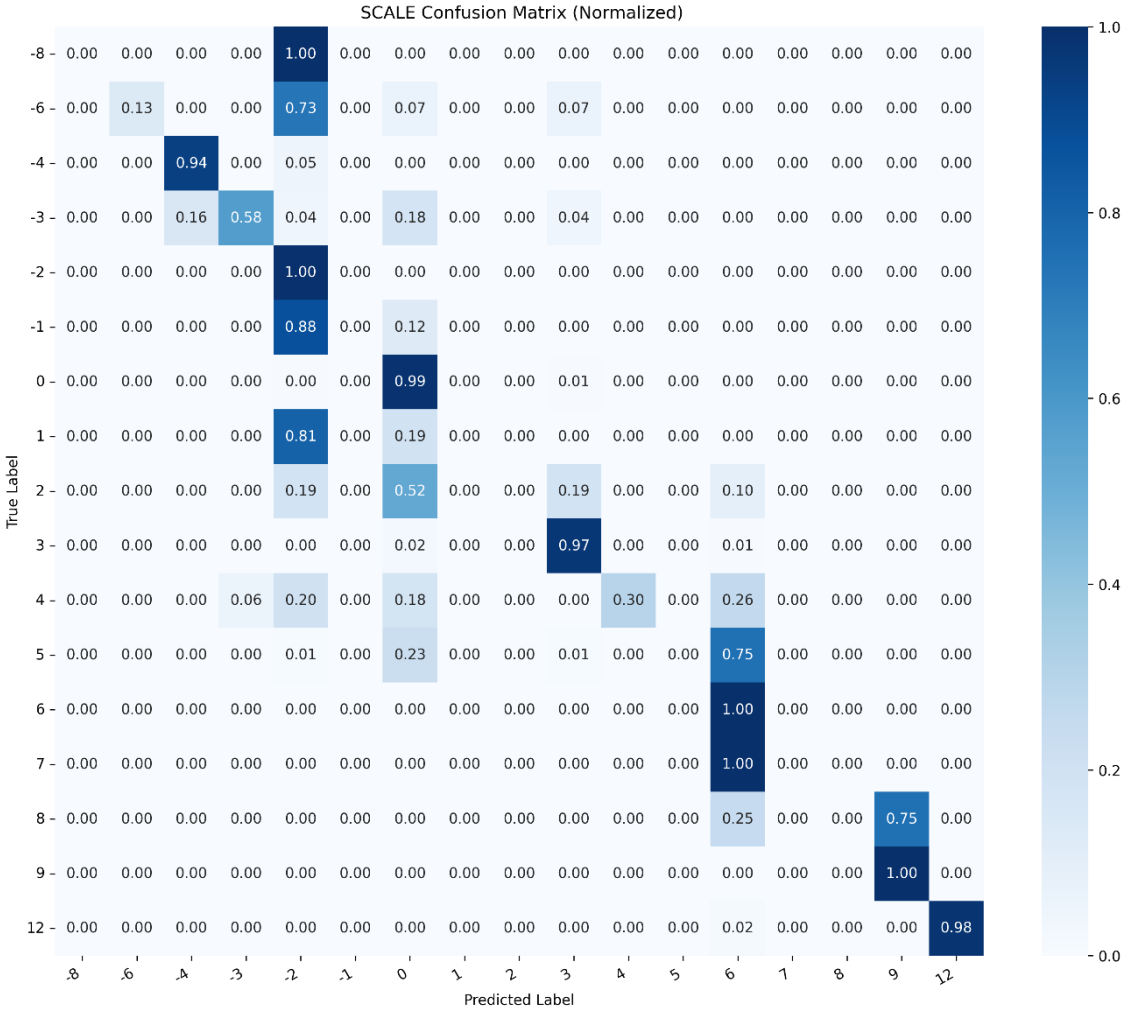



Figure 15 Confusion Matrix of Scale Prediction Result (Normalized)



doi:10.6342/NTU202501600

**Example 1**

context_t The decrease in common stock outstanding at September 30, 2019 was primarily due to the impact of share repurchases of approximately 52 in the nine months ended September 30, 2019 (see Note H).

true scale 6

pred scale 3

Basic earnings per share (EPS) amounts are computed by dividing earnings, after the deduction of preferred stock dividends declared, by the average number of common shares outstanding. Diluted EPS amounts assume the issuance of common stock for all potentially dilutive share equivalents outstanding.

The information used to compute basic and diluted EPS attributable to Arconic common shareholders was as follows (shares in millions):

| | Third quarter ended September 30, | | Nine months ended September 30, | |
|---|---|---|---|---|
| | 2019 | 2018 | 2019 | 2018 |
| Net income | $ 95 | $ 161 | $ 161 | $ 424 |
| Less: preferred stock dividends declared | (1) | (1) | (2) | (2) |
| Net income available to Arconic common shareholders - basic | 94 | 160 | 159 | 422 |
| Add: Interest expense related to convertible notes | 3 | 3 | — | 8 |
| Net income available to Arconic common shareholders - diluted | $ 97 | $ 163 | $ 159 | $ 430 |
| Average shares outstanding - basic | 436 | 483 | 451 | 483 |
| Effect of dilutive securities: | | | | |
| Stock options | 1 | — | 1 | 1 |
| Stock and performance awards | 5 | 5 | 3 | 5 |
| Convertible notes[1] | 15 | 14 | — | 14 |
| Average shares outstanding - diluted | 457 | 502 | 455 | 503 |

[1] The convertible notes matured on October 15, 2019 (see Note Q). No shares of the Company's common stock were issued in connection with the maturity or the final conversion of the convertible notes. As of October 15, 2019, the calculation of average diluted shares outstanding will cease to include the approximately 15 million shares of common stock and the corresponding interest expense previously attributable to the convertible notes.

Common stock outstanding at September 30, 2019 and 2018 was 434 and 483, respectively. The decrease in common stock outstanding at September 30, 2019 was primarily due to the impact of share repurchases of approximately 52 in the nine months ended September 30, 2019 (see Note H). As average shares outstanding are used in the calculation for both basic and diluted EPS, the full impact of share repurchases was not realized in EPS in the third quarter and nine months ended September 30, 2019 as the share repurchases occurred at varying points during 2019.

**Example 2**

context_t As a result, the Company reduced the tax expense related to the impact of the Act to $597 in 2018.

true scale 3

pred scale 6

**BARNES GROUP INC.**
**NOTES TO CONSOLIDATED FINANCIAL STATEMENTS**
**(All dollar amounts included in the notes are stated in thousands except per share data and the tables in Note 22)**

**Example 3**

context_t As of December 31, 2021, the Company had received $10.0 million in prepayments from BASF and applied approximately $0.3 million of credits against amounts invoiced to BASF for Spaceloft A2.

true scale 12

pred scale 6

**Example 4**

context_t The facility is pre-payable at par, after 12 months of call protection (during which prepayment would be at 101% of par), or with respect to prepayments made with respect to a change of control, at 101% of par, and carries a five-year term.

true scale -6

pred scale -2

**Example 5**

context_t In 2019, net adjustments to the estimated profitability of long-term contracts increased AEC operating income by $0.6 million in the first quarter, and $5.0 million in the second quarter.

true scale 7

pred scale 6

Figure 16 Scale Error Example

## 4.4.4 Negative

Lastly, despite the extreme class imbalance in the negative attribute classification, the model did not default to predicting all instances as non-negative. For the minority class (negative = 1), the model achieved a precision of 0.89, recall of 0.74, and an



doi:10.6342/NTU202501600

F1-score of 0.81, indicating that it was able to identify negative values reasonably well despite the skewed distribution. Figure 17 present the confusion matrix of negative attribute.

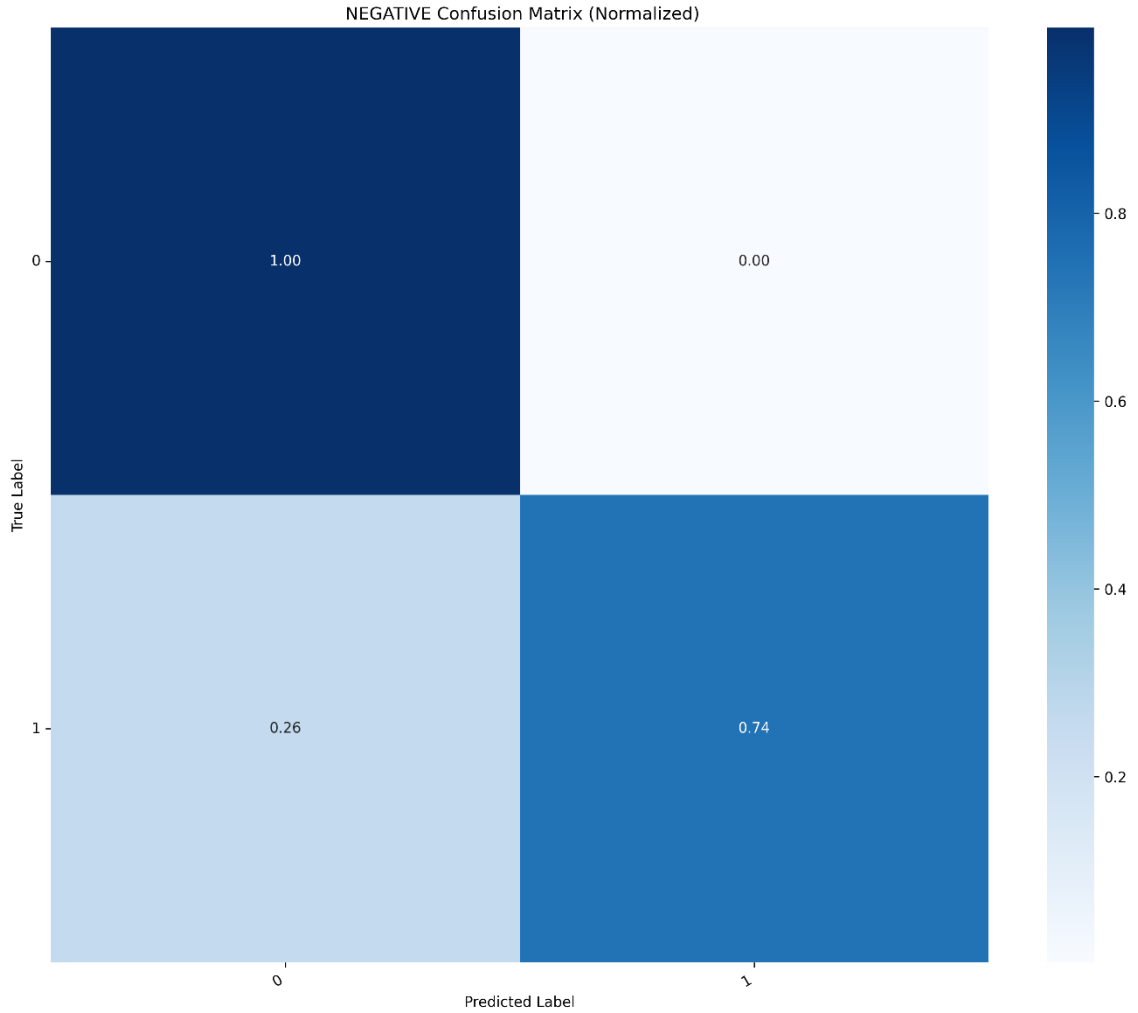



Figure 17 Confusion Matrix of Negative Prediction Result (Normalized)

## 4.5 Managerial Implications

This study offers several important managerial implications for financial reporting and disclosure process:

First, automated XBRL tagging has the potential to significantly reduce manual effort and increase efficiency. By providing a ranked list of recommended tag names for



doi:10.6342/NTU202501600

each fact, the system allows domain experts to choose from a smaller, more relevant set of candidates, rather than searching through the entire taxonomy. Our best model correctly identified the appropriate tag within the top 5 candidates in 97% of cases. Additionally, the system may help reduce reliance on custom concepts by recommending closely aligned standard tags. This can improve the comparability of financial reports across companies and reporting periods.

Second, the proposed framework can enhance the quality and consistency of financial filings. In our experiments, the model was able to identify inconsistencies and potential tagging errors by recognizing patterns across millions of instances. By learning from large-scale historical data, the model reflects how similar facts are typically tagged, serving as both a reference tool and a quality assurance mechanism for human reviewers.

Third, the trained model can be extended to support other financial documents that are not subject to mandatory XBRL requirements, such as shareholder meeting materials, earnings reports, or financial statements published before XBRL became mandatory. This improves the accessibility and usability of unstructured financial information, enabling more structured analysis and bridging the gap between legacy and modern reporting formats.

Fourth, our findings also have important policy implications for regulatory agencies. As shown in Table 35, approximately 84.79% of custom tag names are misclassified, often into existing standard tags. This suggests that many of these custom concepts could potentially be replaced by appropriate standard ones. Given that the SEC has expressed concern over the excessive use of custom concepts, which may hinder comparability across companies (SEC, 2025). our results provide empirical support for



doi:10.6342/NTU202501600

promoting the broader adoption of standard tags. This not only enhances consistency in financial reporting but also aligns with regulatory goals of improving transparency and interoperability of financial disclosures.

Furthermore, our results reveal the current limitations of few-shot learning with LLMs. Despite their flexibility, few-shot LLMs underperform compared to models that are pre-trained and fine-tuned on domain-specific financial texts. This highlights the importance of domain adaptation in financial NLP tasks. By releasing our dataset, we empower companies and researchers to develop and fine-tune their own models for iXBRL automated tagging, tailored to their specific needs and regulatory environments.

These implications suggest that our work is not only feasible for automated XBRL tagging but also valuable for enhancing compliance, improving data quality, and supporting future research across the financial reporting pipeline.



doi:10.6342/NTU202501600

# Chapter 5 Conclusion

There are several research limitations and future directions. First, the current model predicts each target independently, which may ignore the relationship between targets in the same context. Capturing these dependencies may improve overall consistency and accuracy. Second, some attributes are inherently related, such as the tag name often corresponds to a specific time type (e.g., instant or period). However, our current approach does not explicitly model such as dependencies. Third, although we trained the model using all available standard tag name across multiple years, the standard taxonomy may exhibit slight variations over time. Future work could explore how to better handle taxonomy or implement year-specific modeling strategies to adapt to evolving standards. Fourth, this research focuses on the first stage of the task. While later stages for the prediction of more attributes and the generation of custom concept names present further opportunities for exploration. Finally, although few-shot learning has been applied to LLMs in this study, more advanced approaches may yield further improvements in XBRL tagging performance. By addressing these directions, future research can build on this thesis to further enhance the automation and intelligence of XBRL/iXBRL tagging systems in practice.

This thesis presents a study on automatic XBRL tagging, with the following key contributions to the field: First, our work addresses the limitations of previous datasets by constructing a large-scale, fine-grained dataset for automated XBRL tagging. The dataset contains 6.6 million instances and includes a richer set of attributes (such as time, scale, sign, and more), as well as contextual and metadata information. In addition, we develop a data processing pipeline to enhance data quality by addressing common



doi:10.6342/NTU202501600

issues such as cross-page content and missing or inconsistent tags. Second, we reformulate the XBRL tagging task as a multi-attribute prediction problem, which requires jointly predicting multiple attributes for a target text span. This better reflects the real-world complexity of financial tagging. We also design a three-stage task setup to improve the flexibility and usability of the dataset across different modeling and application scenarios. Third, we implement a multi-task learning framework on BERT-based model and apply few-shot prompting using large language models to establish baseline performance. The MTL model demonstrates strong results (e.g., F1 scores of 0.82 for tags, 0.90 for time, 0.99 for scale and sign), showing its ability to handle complex multi-attribute prediction. Our experiments reveal that few-shot LLMs underperform relative to fine-tuned models. This underperformance may stem from the lack of financial pretraining, and the limited capability of basic prompting strategies in handling numeric reasoning and financial semantics. Finally, our dataset and model enable future research in areas such as: tag recommendation systems that recommended 5 to 10 tag name and allow the expert choose from them rather than form whole taxonomy concepts. It also has potential to extend to financial texts that doesn't has XBRL/iXBRL tag, such as shareholder materials or financial filing before mandate of XBRL. Our results reveal the limitations of few-shot prompting using LLM, but the dataset can serve as a foundation for further exploration of LLM-based methods in automated XBRL tagging.



doi:10.6342/NTU202501600

# REFERENCE


Al-Okaily, M., Alkayed, H., & Al-Okaily, A. (2024). Does XBRL adoption increase financial information transparency in digital disclosure environment? Insights from emerging markets. *International Journal of Information Management Data Insights*, *4*(1), 100228. https://doi.org/10.1016/j.jjimei.2024.100228

Amin, K., Eshleman, J. D., & Feng, C. (2018). The Effect of the SEC's XBRL Mandate on Audit Report Lags. *Accounting Horizons*, *32*(1), 1-27.

Bartley, J., Chen, A. Y. S., & Taylor, E. Z. (2011). A Comparison of XBRL Filings to Corporate 10-Ks—Evidence from the Voluntary Filing Program. *Accounting Horizons*, *25*(2), 227-245. https://doi.org/10.2308/acch-10028

Basoglu, K. A., & White, C. E. (2015). Inline XBRL versus XBRL for SEC reporting. *Journal of Emerging Technologies in Accounting*, *12*(1), 189-199.

BLANKESPOOR, E. (2019). The Impact of Information Processing Costs on Firm Disclosure Choice: Evidence from the XBRL Mandate. *Journal of Accounting Research*, *57*(4), 919-967. https://doi.org/10.1111/1475-679X.12268

Caruana, R. (1997). Multitask Learning. *Machine Learning*, *28*(1), 41-75. https://doi.org/10.1023/A:1007379606734

Chang, H. W., Kaszak, S., Kipp, P., & Robertson, J. C. (2021). The Effect of iXBRL Formatted Financial Statements on the Effectiveness of Managers' Decisions When Making Inter-Firm Comparisons. *Journal of Information Systems*, *35*(2), 149-177.

Chouhan, V., & Goswami, S. (2015). XBRL acceptance in India: A behavioral study. *American Journal of trade and Policy*, *2*(2), 71-78.




doi:10.6342/NTU202501600

Chowdhuri, R., Yoon, V. Y., Redmond, R. T., & Etudo, U. O. (2014). Ontology based integration of XBRL filings for financial decision making. *Decision Support Systems*, *68*, 64-76. https://doi.org/10.1016/j.dss.2014.09.004

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., & Brahma, S. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, *25*(70), 1-53.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-Balanced Loss Based on Effective Number of Samples. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* Minneapolis, Minnesota.

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. https://doi.org/10.2308/isys-50399

ESMA. (n.d.). *Electronic Reporting*. European Securities and Markets Authority. https://www.esma.europa.eu/issuer-disclosure/electronic-reporting

Fourny, G. (2023). *The XBRL Book: Simple, Precise, Technical*.

FSA. (2008). *FSA launches new electronic corporate disclosure system (EDINET)*. Financial Services Agency. https://www.fsa.go.jp/en/news/2008/20080317.html

FSA. (2013). *Publication of the next generation EDINET taxonomy*. Financial Services Agency. https://www.fsa.go.jp/search/20130821.html



doi:10.6342/NTU202501600

Gui, L., Leng, J., Zhou, J., Xu, R., & He, Y. (2022). Multi Task Mutual Learning for Joint Sentiment Classification and Topic Detection. *IEEE Transactions on Knowledge and Data Engineering*, *34*(4), 1915-1927. https://doi.org/10.1109/TKDE.2020.2999489

Gupta, S., Chandel, A., & Bhalla, L. (2023). XBRL adoption and information asymmetry: Evidence from the Indian capital market. *Indian Journal of Finance*, *17*(7), 25-36.

Han, S., Kang, H., Jin, B., Liu, X.-Y., & Yang, S. Y. (2024). *XBRL Agent: Leveraging Large Language Models for Financial Report Analysis* Proceedings of the 5th ACM International Conference on AI in Finance, Brooklyn, NY, USA. https://doi.org/10.1145/3677052.3698614

HMRC. (2020). *XBRL guide for businesses*. HM Revenue & Customs. https://www.gov.uk/government/publications/xbrl-guide-for-uk-businesses/xbrl-guide-for-uk-businesses

Hodge, F. D., Kennedy, J. J., & Maines, L. A. (2004). Does Search-Facilitating Technology Improve the Transparency of Financial Reporting? *The accounting review.*, *79*(3), 687-703.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations*, *1*(2), 3.

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.

Janvrin, D. J., & No, W. G. (2012). XBRL Implementation: A Field Investigation to



doi:10.6342/NTU202501600

Identify Research Opportunities. *Journal of Information Systems*, *26*(1), 169-197. https://doi.org/10.2308/isys-10252

Kapil, V., & Martin, D. (2024). *Digital Financial Reporting and AI: Is Automatic XBRL Tagging Feasible Using AI and LLM systems*. UBPartner. https://www.ubpartner.com/digital-financial-reporting-and-ai/

Khatuya, S., Mukherjee, R., Ghosh, A., Hegde, M., Dasgupta, K., Ganguly, N., Ghosh, S., & Goyal, P. (2024, June). Parameter-Efficient Instruction Tuning of Large Language Models For Extreme Financial Numeral Labelling. In K. Duh, H. Gomez, & S. Bethard, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* Mexico City, Mexico.

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2020). Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 318-327. https://doi.org/10.1109/TPAMI.2018.2858826

Liu, C., Luo, X., Sia, C. L., O'Farrell, G., & Teo, H. H. (2014). The impact of XBRL adoption in PR China. *Decision Support Systems*, *59*, 242-249. https://doi.org/10.1016/j.dss.2013.12.003

Liu, C., Luo, X., & Wang, F. L. (2017). An empirical investigation on the impact of XBRL adoption on information asymmetry: Evidence from Europe. *Decision Support Systems*, *93*, 42-50. https://doi.org/10.1016/j.dss.2016.09.004

Liu, C., Wang, T., & Yao, L. J. (2014). XBRL's impact on analyst forecast behavior: An empirical study. *Journal of Accounting and Public Policy*, *33*(1), 69-82. https://doi.org/10.1016/j.jaccpubpol.2013.10.004

Loukas, L., Fergadiotis, M., Chalkidis, I., Spyropoulou, E., Malakasiotis, P.,



doi:10.6342/NTU202501600

Androutsopoulos, I., & Paliouras, G. (2022, May). FiNER: Financial Numeric Entity Recognition for XBRL Tagging. In S. Muresan, P. Nakov, & A. Villavicencio, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* Dublin, Ireland.

Luo, X., Wang, T., Yang, L., Zhao, X., & Zhang, Y. (2023). Initial evidence on the market impact of the iXBRL adoption. *Accounting Horizons*, *37*(1), 143-171.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., & Dokania, P. (2020). Calibrating deep neural networks using focal loss. *Advances in neural information processing systems*, *33*, 15288-15299.

Ni, J., Hernandez Abrego, G., Constant, N., Ma, J., Hall, K., Cer, D., & Yang, Y. (2022, May). Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. In S. Muresan, P. Nakov, & A. Villavicencio, *Findings of the Association for Computational Linguistics: ACL 2022* Dublin, Ireland.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., & Anadkat, S. (2023). Gpt-4 technical report.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002, July). Bleu: a Method for Automatic Evaluation of Machine Translation. In P. Isabelle, E. Charniak, & D. Lin, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* Philadelphia, Pennsylvania, USA.

Ramanan, R. (2024a). *How well do AI models like GPT-4 understand XBRL Data?* XBRL International Inc. https://www.xbrl.org/how-well-do-ai-models-like-gpt-4-understand-xbrl-data/

Ramanan, R. (2024b). *Narrative disclosure analysis with GPT-4*. XBRL International



doi:10.6342/NTU202501600

Inc. https://www.xbrl.org/narrative-disclosure-analysis-with-gpt-4/

Ramanan, R. (2025). *Leveraging LLMs for Smarter Taxonomy Interactions*. XBRL International Inc. https://www.xbrl.org/leveraging-llms-for-smarter-taxonomy-interactions/

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673-2681. https://doi.org/10.1109/78.650093

SEC. (2008). *Interactive Data to Improve Financial Reporting*. Securities and Exchange Commission. https://www.sec.gov/files/rules/final/2009/33-9002.pdf

SEC. (2018). *SEC Adopts Inline XBRL for Tagged Data*. Securities and Exchange Commission. https://www.sec.gov/newsroom/press-releases/2018-117

SEC. (2025). *U.S. GAAP – XBRL Custom Tags Trend for 2022 - 2024*. SEC. https://www.sec.gov/data-research/structured-data/gaap-xbrl-custom-tags

Sharma, S., Khatuya, S., Hegde, M., Shaikh, A., Dasgupta, K., Goyal, P., & Ganguly, N. (2023, July). Financial Numeric Extreme Labelling: A dataset and benchmarking. In A. Rogers, J. Boyd-Graber, & N. Okazaki, *Findings of the Association for Computational Linguistics: ACL 2023* Toronto, Canada.

Taiwan Stock Exchange Corporation. (n.d.). *The eXtensible Business Reporting Language (XBRL)*. Taiwan Stock Exchange Corporation. https://www.twse.com.tw/zh/listed/xbrl.html

Tawiah, V., & Borgi, H. (2022). Impact of XBRL adoption on financial reporting quality: a global evidence. *Accounting Research Journal*, *35*(6), 815-833.



doi:10.6342/NTU202501600

XBRL International. (n.d.). *Release History: XBRL - XBRL Specification*. XBRL International. https://specifications.xbrl.org/release-history-base-spec-xbrl-2.1.html

XBRL International Inc. (2013). *XBRL 2.1*. XBRL International Inc.,. https://specifications.xbrl.org/work-product-index-group-base-spec-base-spec.html

XBRL International Inc. (2024). *AI and XBRL: automatic tagging?* XBRL International Inc. https://www.xbrl.org/news/ai-and-xbrl-automatic-tagging/

XBRL International Inc. (2013). *Inline XBRL 1.1*. XBRL International Inc.,. https://specifications.xbrl.org/work-product-index-inline-xbrl-inline-xbrl-1.1.html

XBRL International Inc. (2019). *iXBRL Tagging Features*. XBRL International Inc. https://www.xbrl.org/guidance/ixbrl-tagging-features/

XBRL US. (2020). *XBRL Taxonomy Development Handbook* XBRL US. https://xbrl.us/xbrl-reference/tdh/

Yang, Y., Uy, M. C. S., & Huang, A. (2020). Finbert: A pretrained language model for financial communications.

Yoon, H., Zo, H., & Ciganek, A. P. (2011). Does XBRL adoption reduce information asymmetry? *Journal of Business Research*, *64*(2), 157-163. https://doi.org/10.1016/j.jbusres.2010.01.008

You, R., Zhang, Z., Wang, Z., Dai, S., Mamitsuka, H., & Zhu, S. (2019). Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. *Advances in neural information processing systems*, *32*.



doi:10.6342/NTU202501600

Yu, J., Marujo, L., Jiang, J., Karuturi, P., & Brendel, W. (2018, oct nov). Improving Multi-label Emotion Classification via Sentiment Classification with Dual Attention Transfer Network. In E. Riloff, D. Chiang, J. Hockenmaier, & J. i. Tsujii, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* Brussels, Belgium.

Zhang, Y., & Shan, Y. (2024). Does The Adoption of Ixbrl Improve Data Usability? Evidence from Future Earnings Response Coefficients (Fercs). *Evidence from Future Earnings Response Coefficients (Fercs)*.

Zhang, Y., & Yang, Q. (2022). A Survey on Multi-Task Learning. *IEEE Transactions on Knowledge and Data Engineering*, *34*(12), 5586-5609. https://doi.org/10.1109/TKDE.2021.3070203

Zhang, Z., Yu, W., Yu, M., Guo, Z., & Jiang, M. (2023, May). A Survey of Multi-task Learning in Natural Language Processing: Regarding Task Relatedness and Training Methods. In A. Vlachos & I. Augenstein, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* Dubrovnik, Croatia.



doi:10.6342/NTU202501600

# APPENDIX

Appendix 1 Document metadata corresponding dei tag

| Attribute | Dei tag |
|---|---|
| company_name | dei:EntityRegistrantName |
| cik | dei:EntityCentralIndexKey |
| document_type | dei:DocumentType |
| period_end_date | dei:DocumentPeriodEndDate |
| fiscal_year | dei:DocumentFiscalYearFocus |
| period_focus | dei:DocumentFiscalPeriodFocus |
| current_fiscal_year_end | dei:CurrentFiscalYearEndDate |

Appendix 2 Standard Prefix of Tag, Axis, and Member, from EDGAR Filer Manual[5]

| Prefix | Taxonomies |
|---|---|
| us-gaap | US GAAP Financial Reporting |
| dei | Document and Entity Information |
| srt | SEC Reporting |
| sic | Standardized Industrial Codes |
| ifrs | International Financial Reporting Standards |
| ebp / us-gaap-ebp | Employee Benefit Plan Financial Reporting |
| cef | Closed-end Fund |
| country | Country (ISO 3166-1) |
| currency | Currency (ISO 4217) |
| cyd | Cybersecurity Disclosure |

[5] https://www.sec.gov/submit-filings/edgar-filer-manual



doi:10.6342/NTU202501600

| | |
|---|---|
| ecd | Executive Compensation Disclosure |
| exch | Exchanges (ISO 10383 MIC) |
| ffd | Filing Fee Disclosures |
| fnd | Funds |
| naics | North American Industry Classification System |
| oef | Open-end Fund |
| rr | Risk/Return (Deprecated as of 2023) |
| rxp | Resource Extraction Payments |
| sbs | Security-Based Swap |
| snj | Subnational Jurisdiction (ISO 3166-2) |
| sro | Self-regulating Organizations |
| stpr | States and Provinces |
| vip | Variable Insurance Products |

## Appendix 3 Standard Prefix of Measure, adapted from Units Registry[6]

| **Prefix** | **Namespace URI** |
|---|---|
| iso4217 | http://www.xbrl.org/2003/iso4217 |
| utr / utreg | http://www.xbrl.org/2009/utr |
| utre | http://www.xbrl.org/2009/utr/errors |
| xbrli | http://www.xbrl.org/2003/instance |

[6] https://www.xbrl.org/specification/utr/rec-2013-11-18/utr-rec-2013-11-18-clean.html



doi:10.6342/NTU202501600

# Appendix 4 Prompt for Few-shot learning

Task:
Predict the attributes of the <target_text> in the following target sentence. The attributes include: “tag”, “time”, “scale”, and “negative”.
- “tag” refers to the corresponding XBRL taxonomy element that best represents the meaning of the <target_text> within the given context.
- “time” represents the reporting period associated with the <target_text>.
  For instant values (i.e., point-in-time data), indicate whether the value refers to the past, current, or future, relative to the report’s end date.
  For period values (i.e., over a duration), choose one of the following based on its relationship to the report’s end date: past, current, future, past_current, past_future, or current_future.
  The classification of past, current, or future is determined with respect to the reporting period, not the actual date of labeling.
  In a 10-K, the reporting period usually spans one full fiscal year ending on the report’s end date.
  In a 10-Q, the reporting period typically covers the most recent three months.
- “scale” is the scaling factor applied to the <target_text> value in the iXBRL format. This represents the order of magnitude of the value: For example: Millions = 6,Thousands = 3, Hundreds = 2, Ones (no scaling) = 0, Hundredths = -2
- “negative” indicates whether the value represented by <target_text> is negative in the context of the XBRL document. true for negative value, false for non-negative.

Example 1:
Input:
context: Proceeds from this offering were used to redeem the €500.0 million of 2.0% Senior Notes due 2023 in March 2021. In connection with this redemption, we paid a “make-whole” amount of $26.2 million (based on the exchange rate of the euro as of the date of redemption) and recognized a loss on extinguishment of $<target_text>28.2<target_text> million, which is included within Other gains and (losses) on our consolidated statements of income.

target text: 28.2
target attributes to predict: ['tag', 'fact', 'time', 'scale']

document information:
company_name: W. P. Carey Inc.
document_type: 10-Q
period_end_date: 2021-06-30
fiscal_year: 2021
period_focus: Q2

Output:
{'tag': 'us-gaap:GainsLossesOnExtinguishmentOfDebt', 'time': 'period; past_current', 'scale': '6', 'negative': True}

// 3 more example

Question:
Input:
context:  For the three and six months ended June 30, 2022, restructuring activities resulted in expenses of $2.6 million and included $1.4 million of severance and other employee benefit costs, $0.9 million of asset-related impairment and accelerated depreciation costs, and $<target_text>0.3<target_text> million of other costs. As of June 30, 2022, $1.3 million of the severance and other employee benefits are unpaid and accrued.

document information:
company_name: ANI PHARMACEUTICALS, INC
document_type: 10-Q
period_end_date: 2022-06-30
fiscal_year: 2022
period_focus: Q2

Output:



doi:10.6342/NTU202501600