XBRLTagRec: Domain-Specific Fine-Tuning and Zero-Shot Re-Ranking with LLMs for Extreme Financial Numeral Labeling

Gang Hu, Qun Zhang, Jingyao Luo, Yile Jiang, Jing Chai, Haiyan Ding

School of Information Science & Engineering, Yunnan University, Kunming, China

***Abstract*—Publicly traded companies must disclose financial information under regulations of the Securities and Exchange Commission (SEC) and the Generally Accepted Accounting Principles (GAAP). The eXtensible Business Reporting Language (XBRL), as an XML-based financial language, enables standardized and machine-readable reporting, but accurate tag selection from large taxonomies remains challenging. Existing fine-tuning-based methods struggle to distinguish highly similar XBRL tags, limiting performance in financial data matching. To address these issues, we introduce XBRLTagRec[1], an end-to-end framework for automated financial numeral tagging. The framework generates semantic tag documents with a fine-tuned FLAN-T5-Large model, retrieves relevant candidates via semantic similarity, and applies zero-shot re-ranking with ChatGPT-3.5 to select the optimal tag. Experiments on the FNXL dataset show that XBRLTagRec outperforms the state-of-the-art FLAN-FinXC framework, achieving 2.64%–4.47% improvements in Hits@1 and Macro metrics. These results demonstrate its effectiveness in large-scale and semantically complex tag matching scenarios.**

***Index Terms*—Financial Numeral Labelling, Large Language Model, Zero-Shot Re-Ranking, XBRL**

I. Introduction

Publicly listed companies are required to disclose standardized financial information in compliance with regulatory frameworks like GAAP [1]. XBRL, an XML-based standard, is widely adopted to enable structured, machine-readable reporting, improving the interpretability and comparability of financial data [2]. Accurately assigning concept tags in complex financial texts remains a key challenge.

In practice, XBRL tag selection remains largely manual, requiring substantial domain expertise and contextual understanding [3]. The large scale of XBRL taxonomies, semantic overlap among tag definitions, and inconsistent financial language usage make automated tag matching highly challenging. These issues lead to frequent annotation errors and limit the scalability of intelligent financial data processing.

Prior work has explored automated financial numeral labeling from multiple perspectives. FiNER [4] formulated XBRL tag assignment as a NER task using BERT-based sequence labeling, but its limited tag coverage restricts real-world applicability. Subsequent studies applied extreme multi-label classification on the FNXL dataset [5], yet such supervised methods struggle to distinguish highly similar tags and fail to fully exploit the semantic structure of XBRL taxonomies. More recent approaches incorporate tag semantics via graph-based models or generative frameworks [6]–[8], while fine-tuned financial LLMs mainly focus on entity-centric NER rather than financial numerals [9], [10]. Consequently, accurate semantic differentiation and ranking of extremely similar XBRL numeral tags remains an open challenge.

Despite the progress made by these methods from various perspectives, they still face three main challenges in practical applications: (1) In ultra-large tag systems, NER and classification methods struggle to effectively differentiate semantically similar tags; (2) Existing methods generally rely on static tag encoding, which fails to fully leverage the rich semantic features of XBRL tag documents; (3) While generative methods can provide candidate label documents, they still lack robust and efficient semantic ranking mechanisms to filter the optimal results. These issues collectively lead to performance bottlenecks in current models. As shown in Figure 1, FLAN-FinXC is an effective state-of-the-art baseline for financial numeral labeling. However, it may still miss the correct label within its Top-k matches when relying on Top-1 matching. In contrast, the introduction of ChatGPT, with its broad financial knowledge and strong learning capabilities, enables iterative differentiation of these highly challenging, extremely similar labels, leading to more accurate label predictions.

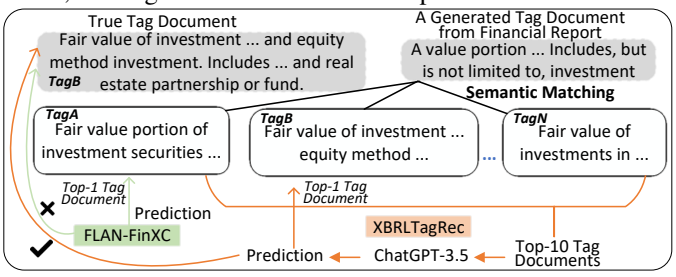

Fig. 1. Illustrates the effectiveness of ChatGPT-3.5 in semantic re-ranking extremely similar label documents.

In this paper, we propose XBRLTagRec, an end-to-end framework for automated financial numeral labeling that matches target numbers in financial texts to standardized XBRL tags. XBRLTagRec leverages instruction-tuned LLMs to generate label documents and integrates semantic retrieval with zero-shot re-ranking to select the optimal tag from a large candidate set, improving both accuracy and interpretability. XBRLTagRec consists of three modules: (1) Tag Document

*Corresponding Author: dinghaiyan@ynu.edu.cn

[1]https://github.com/ckw29/XBRLTagRec

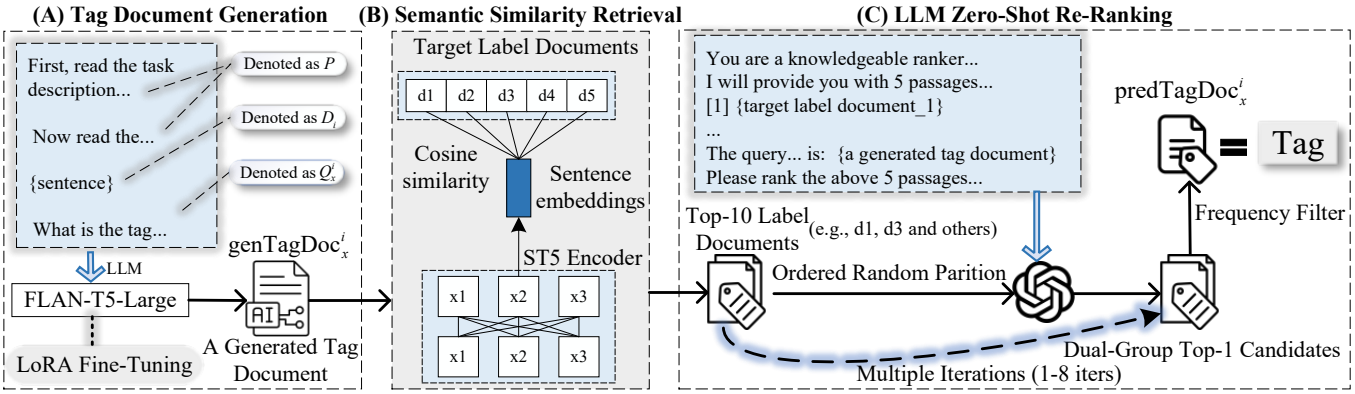

Fig. 2. Our proposed XBRLTagRec framework consists of three components: (A) Tag Document Generation, (B) Semantic Similarity Retrieval, and (C) LLM Zero-Shot Re-Ranking. First, a FLAN-T5 LoRA-tuned LLM generates a tag document from instruction prompts, financial text, and related numeric questions. Next, the Sentence-T5-XXL model encodes this document and all ground-truth XBRL documents to retrieve the Top-10 most similar ones via cosine similarity. Finally, ChatGPT-3.5 performs multi-round re-ranking of these candidates, and the final XBRL tags are selected by majority vote.

Generation, where FLAN-T5-Large is fine-tuned with LoRA to generate semantic label documents from financial text, prompts, and target numbers; (2) Semantic Similarity Retrieval, which uses Sentence-T5-XXL to retrieve the Top-10 most relevant XBRL label documents; (3) LLM Zero-Shot Re-Ranking, where ChatGPT performs multi-round semantic re-ranking with frequency voting to determine the final label. This design forms a closed-loop pipeline from generation to selection and effectively exploits the rich semantics of XBRL tag documents in large-scale, highly similar label spaces.

To evaluate effectiveness, we conduct experiments on a subset of the FNXL dataset. XBRLTagRec outperforms the state-of-the-art baseline FLAN-FinXC, achieving improvements of 2.64%–4.47% in Hits@1 and Macro metrics, showing strong robustness in accurate prediction and multi-label balance.

In summary, the main contributions are as follows: **(1) We propose an XBRL tag prediction framework, XBRLTagRec**, which combines the generative capabilities of LLMs, semantic similarity retrieval, and multi-round iterative re-ranking mechanism. It builds an end-to-end pipeline solution that enhances the accuracy of matching financial text numbers with XBRL tags. **(2) We introduce an iterative re-ranking method via filtering and frequency voting**, leveraging the semantic reasoning capabilities and statistical properties of ChatGPT-3.5. This approach achieves robust ranking of candidate label documents, significantly improving the stability and semantic alignment of predicted labels. **(3) We develop a modular and highly pluggable framework architecture**, with clear boundaries between each module. This modular structure allows for flexible replacement of generators, encoders, and re-rankers, thus supporting the seamless integration of different LLMs and inference strategies. This flexibility enhances the system's generalizability and engineering adaptability.

II. RELATED WORK

(1) XBRL Tagging Methods: To improve the accuracy of XBRL tag annotation for numerical data in financial texts, early studies used structured knowledge and domain-specific modeling techniques. [6] proposed GalaXC, which built a graph neural network based on hierarchical relationships among tags to enhance the model's understanding of semantic structures. [5] released the FNXL dataset and used a pipeline-based approach. [11] integrated various semantic similarity metrics to align custom tags with standard taxonomies. With the rapid progress of LLMs, researchers began to explore their potential in financial tagging tasks. [8] proposed the FLAN-FinXC framework, which reformulated the extreme multi-label tagging problem as a generative task. [12] developed the XBRL-Agent system to enhance LLMs' comprehension of financial reports. **(2) Re-Ranking with LLMs:** Recently, LLM-based re-ranking methods have achieved strong results in cross-lingual retrieval, multi-passage ranking, and contextual modeling. For example, [13] proposed the unsupervised list-wise method LRL, improving cross-lingual retrieval without fine-tuning. [14] used fine-tuned LLMs to re-rank paragraph-level texts by relevance. [15] introduced RankVicuna for zero-shot listwise document re-ranking. [16] proposed the Pairwise Ranking Prompting (PRP) method to enhance LLM ranking via pairwise comparison. [17] presented the ICR method, which uses attention signals for contextual re-ranking and eases the efficiency bottleneck of generative models.

III. METHODOLOGY

In this section, we present the XBRLTagRec framework for predicting XBRL numeral tags, as illustrated in Figure 2. Descriptions of these three components are provided below.

A. Tag Document Generation

We adopt the instruction-tuned FLAN-T5-Large model [18] as our core generator for text generation under an instruction learning paradigm. In this process, we use parameter-efficient LoRA [19] fine-tuning to FLAN-T5-Large using the training set, significantly reducing training costs while enhancing the model's adaptability to specific financial tasks. Each training sample consists of three components: an instruction prompt (denoted as P), a financial statement text D_i, and a question Q_x^i targeting a specific numeral in the statement. These three components are concatenated to form the final input:

$$\mathrm{D}_x^i = P \oplus D_i \oplus Q_x^i \tag{1}$$

where i denotes the i-th financial report, x denotes a specific numeral within the financial statement text D_i, and $\oplus$ denotes a newline-based joining operation.

The objective of this process is to train the model to generate the corresponding XBRL tag document for the target numeral, marked as genTagDoc_x^i, where LLM_{LoRA} denotes the FLAN-T5-Large model fine-tuned with the LoRA technique.

$$\text{genTagDoc}_x^i = \text{LLM}_{\text{LoRA}}(D_x^i) \tag{2}$$

During training, we adopt an autoregressive approach to generate the tag document step by step, using the cross-entropy loss as the optimization objective. At each time step, the model generates the next token based on the current input state until the tag document is produced. After training, the fine-tuned model is used during the testing phase, where the input format remains the same as in the training phase, and the output is the generated tag document genTagDoc_x^i.

It is important that we generate XBRL tag documents rather than individual XBRL tags. Although there exists a one-to-one correspondence between the two, tag documents provide richer semantic structures, enabling finer-grained semantic distinctions and enhancing the model's discriminative ability in extreme classification tasks. Meanwhile, the introduction of the LoRA fine-tuning mechanism avoids full-parameter updates and greatly reduces resource use during training.

B. Semantic Similarity Retrieval

We adopt the pretrained model [20] Sentence-T5-XXL [21] as the embedding encoder to encode each generated tag document genTagDoc_x^i produced by the LoRA-tuned FLAN-T5-Large LLM. Similarly, all ground-truth XBRL tag documents are also encoded into their corresponding embedding vectors. After obtaining all embedding vectors, we compute the cosine similarity between genTagDoc_x^i and each ground-truth XBRL tag document embedding. We then return the Top-10 XBRL tag document set with the highest similarity scores.

C. LLM Zero-Shot Re-Ranking

To further improve the accuracy of tag prediction, we introduce the LLM Zero-Shot Re-Ranking mechanism. We employ ChatGPT-3.5 (API version is gpt-3.5-turbo-1106) as a re-ranker to refine the Top-10 candidate XBRL tag documents obtained in the previous stage with Sentence-T5-XXL. We select the XBRL tag document that is semantically closest to genTagDoc_x^i as the final predicted document predTagDoc_x^i, and thus derive the predicted XBRL tag. To fully exploit the semantic information in candidate tag documents and mitigate local biases in the ranking process, we design a multi-round iterative re-ranking strategy that combines *filtering* and *most-frequent selection*. The detailed procedure is as follows:

(a) Filtering Strategy: First, we randomly shuffle the Top-10 candidate documents while preserving their original similarity-based ranking order within each group. The shuffled candidates are equally split into two groups, each containing 5 documents. Next, the target XBRL tag documents of each group and the generated tag document genTagDoc_x^i are jointly fed into the ChatGPT-3.5 re-ranking module [14]. This module re-ranks the target tag documents based on their relevance to genTagDoc_x^i. The target XBRL tag document with the highest relevance in the ranking results is then returned as the most relevant candidate in the group.

(b) Multi-Round Iteration: We repeat the filtering process for several rounds. In each round, one optimal document is selected from each of the two groups, and the selected documents are added to a candidate pool for final selection. Since each iteration adopts a different combination of candidates, this mechanism effectively enhances semantic diversity and mitigates local biases in the ranking process, thereby improving the re-ranking module's robustness and global coverage.

(c) Most-Frequent Selection: We count the cumulative frequency with which each document is selected as the top-ranked candidate across all iterations, denoted as $f(d)$, whose computation is defined in Eq.3:

$$\text{f(d)} = \sum_{t=1}^{T} \mathbb{I}[d = d_t^*] \tag{3}$$

where T denotes the number of iterations, d_t^* denotes the top-ranked document in iteration t, and $\mathbb{I}[\cdot]$ is an indicator function.

Finally, we select the document with the highest frequency $f(d)$ as the final tag document prediction predTagDoc_x^i for the generated tag document genTagDoc_x^i. In addition, during implementation, we design a fixed input format with strong instruction constraints for ChatGPT-3.5 to ensure that it returns rankings in a consistent and structured manner. This prompt is adapted from [14], which is selected due to its strong performance in prior experiments. Based on this, we further refine the original prompt to develop the high-quality re-ranking prompt used in this study. The details are shown in Figure 3. Overall, the complete algorithmic workflow of the XBRLTagRec framework is shown in Algorithm 1.

> You are a knowledgeable ranker, an intelligent assistant that can rank passages based on their relevancy to the query.I will provide you with 5 passages, each indicated by number identifier **[ ]**. Rank the passages based on their relevance to query: **{a generated tag document}** \n
> [1] **{target label document_1}** ⋯ [5] **{target label document_5}** \n
> The query for this search is: **{a generated tag document}**
> Please rank the above 5 passages... Strictly use the following output format to return the five passage numbers: [number1] > [number2] > [number3] > [number4] > [number5]. Do not ... adding any extra words or explanations.

Fig. 3. The prompt guides the LLM to rank the five target label documents by their relevance to the generated tag document and the target label documents.

IV. EXPERIMENTAL SETUP

A. Dataset

The FNXL dataset was constructed by [5] based on publicly available SEC annual 10-K filings. It covers 2,339 publicly listed companies and contains 79,088 sentences along with 142,922 manually annotated financial numeral instances. Each numeral is associated with a corresponding GAAP metric and assigned an appropriate XBRL tag, resulting in a total of 2,794 distinct tags. Notably, the tag distribution exhibits a pronounced long-tail characteristic. *Since the full FNXL dataset [5] is not publicly available, we conduct experiments*

on the partial FNXL dataset released by Khatuya et al. [8]. This subset is derived from the same original dataset and represents the publicly accessible portion used in our experiments.

Algorithm 1 The process for XBRLTagRec framework

Require: (1) Financial report: D_i, $i \in [1, M]$; (2) Instruction prompt: P; (3) Target Numeral: x; (4) Question: Q_x^i, $i \in [1, M]$ (5) Ground-truth XBRL tag documents: trueTagDoc$_x^i$, $i \in [1, M]$
Ensure: The predicted XBRL tag document: predTagDoc$_x^i$;
  for $i = 1; i \leq M; i++$ **do**
    $D_x^i = P \oplus D_i \oplus Q_x^i$, where $\oplus$ denotes newline concatenation;
  end for
  for $i = 1; i \leq M; i++$ **do**
    Generates tag documents: genTagDoc$_x^i$ = LLM$_{\text{LoRA}}(D_x^i)$;
  end for
  for $i = 1; i \leq M; i++$ **do**
    Compute embeddings vectors: $\mathbf{Vgen}_x^i$ = Enc(genTagDoc$_x^i$), $\mathbf{Vtrue}_x^i$ = Enc(trueTagDoc$_x^i$);
  end for
  Top10_Docs = TopK$_{k=10}$ $\left(\cos(Vgen_x^i, Vtrue_x^i)\right)$
  Initialize vote counter f : Top10_Docs $\to \mathbb{N}_0$
  for $t = 1; t \leq T; t++$ **do**
    Randomly divide Top-10 tag documents into two groups G_1 and G_2, each with 5 documents, *while preserving intra-group ranking order*
    for $k = 1; k \leq 2; k++$ **do**
      Input G_k and genTagDoc$_x^i$ into ChatGPT-3.5 to obtain ranking
      Select top-ranked tag document d_t^k
    end for
    Update frequency count: $f(d_t^k) \leftarrow f(d_t^k) + 1$, $k \in [1, 2]$
  end for
  Select the document with the highest vote count: predTagDoc$_x^i$ = $\arg\max_d f(d)$

B. Evaluation Metrics

We employ four evaluation metrics for XBRL tag prediction. **Macro-Precision** (M-P): computes precision for each class and averages the results. **Macro-Recall** (M-R): averages recall across all classes. **Macro-F1** (M-F1): balances precision and recall and is suitable for class-imbalance scenarios [22]. In addition, to reflect practical utility and support expert tag recommendation, we introduce **Hits@1** [23].

C. Baseline Methods

To confirm the effectiveness of our framework, we compare its performance against the following baselines: **(1) FLAN-FinXC** [8]: A strong baseline method that achieves notable performance on extreme financial number labeling tasks. It is built on a generative model and optimized through instruction tuning. **(2) AttentionXML Pipeline** [5]: This method builds a BERT-based sequence-to-sequence tagger and models the tagging task as an extreme multi-label classification problem. **(3) FiNER** [4]: This method formulates the tagging task as an NER problem, but only covers the 139 most frequent XBRL tags. **(4) Label Semantics** [7]: This method incorporates semantic descriptions of labels and reformulates the tagging task as a standard NER task with entity descriptions as input. **(5) GalaXC** [6]: This method performs joint learning over a constructed document–label graph, effectively integrating label meta-information into the model's representation.

D. Implementation

We implement our framework on a 32GB Tesla V100 GPU and use the pretrained checkpoints of FLAN-T5-Large and Sentence-T5-XXL from Hugging Face. During the LoRA fine-tuning process based on the FLAN-T5-Large model, we set the rank to 2, the scaling factor (lora_alpha) to 32, the dropout rate to 0.05, the learning rate to 5e-4, the number of training epochs to 5, the maximum length of input sequence to 128, and batch size to 8. In the re-ranking module using ChatGPT-Turbo-1106, we divide the Top-10 candidate XBRL tag documents into two groups, each containing 5 documents (group size = 5). In each iteration, the top-ranked tag document from each group is selected and added to a candidate pool for final filtering. This iterative procedure is repeated for 13 rounds.

E. Research Questions

RQ1: Does our proposed XBRLTagRec framework outperform baseline methods?

RQ2: To what extent do the key parameter designs of our framework affect its overall performance?

RQ3: Does our framework demonstrate generalization and robust performance across various LLMs?

V. RESULTS

A. *RQ1: Does our proposed XBRLTagRec framework outperform baseline methods?*

To address RQ1, we conduct comparative evaluations of several existing methods and the proposed framework XBRLTagRec on the FNXL dataset. The results are presented in Table I. XBRLTagRec achieves the best performance across all evaluation metrics, namely Hits@1, M-P, M-R, and M-F1, highlighting its advantage in the XBRL tagging task.

TABLE I
PERFORMANCE EVALUATION BASED ON MACRO AND HITS@1 METRICS. THE BEST RESULT IS SHOWN IN BOLD.

Method	Eval Metrics			
	Hits@1	M-P	M-R	M-F1
GalaXC	0.3596	0.2088	0.2193	0.1870
AttentionXML	0.7676	0.5069	0.4851	0.4754
Label Semantics	0.5815	0.3360	0.3135	0.3058
FiNER	0.6736	0.4104	0.4140	0.3896
FLAN-FinXC	0.8354	0.4724	0.4881	0.4616
XBRLTagRec$_{\text{(1-2 iters)}}$	0.8526	0.4936	0.5018	0.4777
XBRLTagRec$_{\text{(1-8 iters)}}$	**0.8618**	**0.5171**	**0.5323**	**0.5050**

Specifically, compared with existing baselines, XBRLTagRec already surpasses all baseline methods when the iteration range is limited to 1–2 rounds. When the range is extended to 1–8 rounds, the performance of XBRLTagRec further improves, with Hits@1 increasing to 0.8618. Relative to the strongest baseline, FLAN-FinXC, Hits@1, M-P, M-R, and M-F1 improve by 2.64%, 4.47%, 4.42%, and 4.34%, respectively. These improvements demonstrate that XBRLTagRec not only achieves higher accuracy in top-ranked label prediction but also attains a more balanced enhancement between precision and recall, leading to a substantial improvement in overall F1 performance. In comparison with other baselines such as FiNER, the improvements are even more pronounced; for example, relative to FiNER, Hits@1 increases by 18.82% and M-F1 by 11.54%, further validating the effectiveness of the proposed framework for XBRL tagging. In summary,

XBRLTagRec outperforms all baselines on the FNXL dataset, underscoring its strong practical value and potential for adoption in financial XBRL tagging tasks.

Answer for RQ1: *Our XBRLTagRec framework outperforms all baselines in the XBRL tagging task within the financial domain.*

B. *RQ2: To what extent do the key parameter designs of our framework affect its overall performance?*

To answer RQ2, we conduct three empirical studies focusing on three core parameters: the iteration range, the group size, and the group ordering strategy. The results are presented in Table II-Table IV, respectively. We then perform a comparative analysis based on the evaluation metrics.

(1) Impact of different iteration ranges on model performance: Table II shows the tag prediction performance of XBRLTagRec under different iteration settings. The performance shows an "initial increase followed by stability" trend as iterations increase. Hits@1 reaches its highest value at iterations 1-13, while M-P, M-R, and M-F1 achieve their best values at iterations 1-8. The average values of the four metrics (denoted as avg) indicate that iterations 1-8 provide the optimal overall performance, balancing performance improvement with computational cost. After the 8th iteration, the performance improvement slows, and metrics such as M-F1 decline. This suggests that excessive iterations may introduce redundant information, reducing the discriminative power of the re-ranking process. Therefore, selecting an appropriate number of iterations can enhance performance, prevent overfitting, and avoid unnecessary computation. The results confirm that the iteration selection mechanism of XBRLTagRec is stable, convergent, and scalable under reasonable settings.

TABLE II
THE IMPACT OF ITERATION RANGE ON XBRLTAGREC'S RE-RANKING MODULE. THE BEST RESULT IS SHOWN IN BOLD.

Iteration Range	Eval Metrics				
	Hits@1	M-P	M-R	M-F1	AVG
1-2 iters	0.8526	0.4936	0.5018	0.4777	0.5814
1-3 iters	0.8543	0.5028	0.5140	0.4887	0.5900
1-4 iters	0.8581	0.5089	0.5193	0.4947	0.5953
1-5 iters	0.8599	0.5112	0.5274	0.5002	0.5997
1-6 iters	0.8619	0.5157	0.5305	0.5039	0.6030
1-7 iters	0.8624	0.5146	0.5304	0.5032	0.6027
1-8 iters	0.8618	**0.5171**	**0.5323**	**0.5050**	**0.6041**
1-9 iters	0.8634	0.5110	0.5300	0.5018	0.6016
1-10 iters	0.8634	0.5126	0.5312	0.5026	0.6025
1-11 iters	0.8634	0.5114	0.5315	0.5025	0.6022
1-12 iters	0.8636	0.5097	0.5299	0.5010	0.6011
1-13 iters	**0.8640**	0.5099	0.5287	0.5005	0.6008

(2) Impact of different group sizes on model performance: Table III reports the results of XBRLTagRec on the FNXL dataset under different group size settings. The experimental results show that group size has a significant impact on performance: when it increases from 3 to 5, all evaluation metrics improve, indicating that a larger candidate group can enhance the model's discriminative capability and decision space. However, when the group size is further expanded to 6 or 7, although Hits@1 increases slightly, metrics such as M-F1 exhibit degradation and the average score no longer improves. This suggests that an excessively large group size yields only marginal gains and may even introduce distracting documents. Therefore, setting the group size to 5 achieves the best balance between performance and computational cost, further confirming that an appropriate group size in the re-ranking stage is a key factor influencing overall effectiveness.

TABLE III
THE IMPACT OF GROUP SIZE ON THE XBRLTAGREC RE-RANKING MODULE. THE BEST RESULT IS SHOWN IN BOLD.

Group Size	Eval Metrics				
	Hits@1	M-P	M-R	M-F1	AVG
3	0.8500	0.5023	0.5167	0.4903	0.5898
4	0.8601	0.5056	0.5209	0.4944	0.5953
5	0.8618	**0.5171**	**0.5323**	**0.5050**	**0.6041**
6	**0.8783**	0.5116	0.5262	0.4997	0.6040
7	0.8733	0.5117	0.5263	0.4985	0.6025

(3) Impact of different grouping strategies on model performance: Table IV presents the performance of XBRLTagRec under two grouping strategies: Order-Preserving, which maintains the original cosine similarity ranking, and Order-Shuffled, which applies random sorting. The results indicate that Order-Preserving significantly outperforms Order-Shuffled, achieving higher scores across all metrics. For instance, Order-Shuffled drops to 0.4663 in M-F1, and avg decreases by approximately 3%. This suggests that keeping the cosine similarity ranking helps the model make more accurate re-ranking decisions, while random sorting reduces the effectiveness of re-ranking. These experimental results demonstrate that properly preserving the initial semantic ranking contributes to improving ranking accuracy and robustness.

TABLE IV
THE IMPACT OF GROUPING STRATEGIES IN THE XBRLTAGREC'S RE-RANKING MODULE. THE BEST RESULT IS HIGHLIGHTED IN BOLD.

Grouping Strategy	Eval Metrics				
	Hits@1	M-P	M-R	M-F1	AVG
Order-Preserving	**0.8618**	**0.5171**	**0.5323**	**0.5050**	**0.6041**
Order-Shuffled	0.8557	0.4900	0.4871	0.4663	0.5748

In summary, to use the re-ranking module, we adopt this configuration for XBRLTagRec: 1–8 iterations to balance performance and cost, a group size of 5 to control ranking interference, and an Order-Preserving strategy to keep cosine-based ranking consistency. This setting yields stable performance across experiments and offers guidance for future tasks.

Answer for RQ2: *Three key parameters—iteration range, group size, and grouping strategy—are crucial for performance gains.*

C. *RQ3: Does our framework demonstrate generalization and robust performance across various LLMs?*

To answer RQ3, we replace the LLM in the re-ranking module under a unified generation and retrieval pipeline, selecting four mainstream large-scale models—ChatGPT-3.5-Turbo-1106 [24], GPT-4o [24], DeepSeek-V3 [25], and ERNIE-3.5-8K [26]—as re-rankers to evaluate their performance on the FNXL dataset. The detailed results are presented in Table V.

As shown in Table V, the XBRLTagRec framework maintains stable and strong performance across all four models, demonstrating good model-independence and cross-model

TABLE V
PERFORMANCE OF XBRLTAGREC WITH DIFFERENT LLM-BASED RE-RANKERS (EVALUATED WITH 1–2 ITERATIONS)

Model+LLM	Eval Metrics				
	Hits@1	M-P	M-R	M-F1	AVG
FLAN-FinXC (SOTA)	0.8354	0.4724	0.4881	0.4616	0.5644
$\text{XBRLTagRec}_{+\text{ChatGPT-3.5}}$	0.8526	0.4936	0.5018	0.4777	0.5814
$\text{XBRLTagRec}_{+\text{DeepSeek-V3}}$	0.8672	0.4856	0.4935	0.4706	0.5792
$\text{XBRLTagRec}_{+\text{ERNIE-3.5-8K}}$	0.8594	0.4992	0.5028	0.4827	0.5860
$\text{XBRLTagRec}_{+\text{GPT-4}}$	0.8730	0.5160	0.5227	0.4998	0.6029

adaptability. Although discrepancies still exist among the models, the overall variance remains within a controllable range, indicating that the framework exhibits strong robustness under model-replacement scenarios. Specifically, GPT-4 achieves the best performance across all metrics; ERNIE-3.5-8K ranks second to GPT-4 in terms of M-F1, making it more suitable for scenarios that require a balance between precision and recall; DeepSeek-V3 attains a higher Hits@1 but the lowest M-F1, reflecting its bias toward top-ranking optimization, although ChatGPT-3.5 has the lowest Hits@1, its M-F1 exceeds that of DeepSeek-V3, indicating a more stable ranking capability. Considering both overall performance and computational cost, this study ultimately selects ChatGPT-3.5 as the primary experimental model. In summary, the experiment keeps the core structure of the framework unchanged and only replaces the LLM in the re-ranking module. The results show that the framework produces stable and reliable outcomes across different models, highlighting its modular and plug-and-play design advantages, as well as demonstrating good scalability that supports convenient integration of more types of LLMs.

Answer for RQ3: *Considering computational cost, XBRLTagRec outperforms baselines with only two iterations across LLMs.*

VI. CONCLUSION

In conclusion, we propose XBRLTagRec, an end-to-end framework for automated XBRL tag matching in financial texts. By integrating instruction-tuned LLMs, semantic retrieval, and iterative re-ranking, XBRLTagRec effectively addresses semantic ambiguity and contextual complexity in large-scale tag systems. Experimental results demonstrate clear improvements over existing methods, including the state-of-the-art FLAN-FinXC. Its modular design further ensures flexibility and scalability, making XBRLTagRec a practical and robust solution for financial data annotation.

VII. LIMITATIONS

This study has three limitations: (1) iterative re-ranking with ChatGPT-3.5 increases computational and API costs; (2) the impact of different label embedding models is not examined; (3) zero-shot numeral labeling using untuned general LLMs is not investigated. Future work will explore efficient zero-shot methods and external knowledge integration.

ACKNOWLEDGMENTS

The authors thank the General Program of Applied Basic Research of Yunnan Province (No.202301AT070184), the Open Project Program of Yunnan Key Laboratory of Intelligent Systems and Computing (No.ISC22Y08), and the Open Research Project of the Yunnan University (YNU) Resilience and Excellence Children's Character Development Platform (NO.K207003250006).

REFERENCES

[1] L. Enriques, S. Gilotta *et al.*, "Disclosure and financial market regulation," *The Oxford handbook of financial regulation*, pp. 511–36, 2015.
[2] J. Richards, B. Smith, and A. Saeedi, "An introduction to xbrl," *Available at SSRN 1007570*, 2006.
[3] I. Troshani and A. Lymer, "Translation in xbrl standardization," *Information Technology & People*, vol. 23, no. 2, pp. 136–164, 2010.
[4] L. Loukas, M. Fergadiotis *et al.*, "Finer: Financial numeric entity recognition for xbrl tagging," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 4419–4431.
[5] S. Sharma *et al.*, "Financial numeric extreme labelling: A dataset and benchmarking," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 3550–3561.
[6] D. Saini, A. K. Jain, K. Dave *et al.*, "Galaxc: Graph neural networks with labelwise attention for extreme classification," in *Proceedings of the Web Conference 2021*, 2021, pp. 3733–3744.
[7] J. Ma, M. Ballesteros *et al.*, "Label semantics for few shot named entity recognition," *arXiv preprint arXiv:2203.08985*, 2022.
[8] S. Khatuya, R. Mukherjee, A. Ghosh *et al.*, "Parameter-efficient instruction tuning of large language models for extreme financial numeral labelling," *arXiv preprint arXiv:2405.06671*, 2024.
[9] G. Hu, K. Qin, C. Yuan *et al.*, "No language is an island: Unifying chinese and english in financial large language models, instruction data, and benchmarks," *arXiv preprint arXiv:2403.06249*, 2024.
[10] Y. Wang, Y. Ren *et al.*, "Fintagging: An llm-ready benchmark for extracting and structuring financial information," *arXiv preprint arXiv:2505.20650*, 2025.
[11] R. Wang, "Standardizing xbrl financial reporting tags with natural language processing," *Available at SSRN 4613085*, 2023.
[12] S. Han, H. Kang *et al.*, "Xbrl agent: Leveraging large language models for financial report analysis," in *Proceedings of the 5th ACM International Conference on AI in Finance*, 2024, pp. 856–864.
[13] X. Ma, X. Zhang *et al.*, "Zero-shot listwise document reranking with a large language model," *arXiv preprint arXiv:2305.02156*, 2023.
[14] W. Sun, L. Yan, X. Ma *et al.*, "Is chatgpt good at search? investigating large language models as re-ranking agents," *arXiv preprint arXiv:2304.09542*, 2023.
[15] R. Pradeep, S. Sharifymoghaddam, and J. Lin, "Rankvicuna: Zero-shot listwise document reranking with open-source large language models," *arXiv preprint arXiv:2309.15088*, 2023.
[16] Z. Qin, R. Jagerman *et al.*, "Large language models are effective text rankers with pairwise ranking prompting," *arXiv preprint arXiv:2306.17563*, 2023.
[17] S. Chen, B. J. Gutiérrez *et al.*, "Attention in large language models yields efficient zero-shot re-rankers," *arXiv preprint arXiv:2410.02642*, 2024.
[18] H. W. Chung *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.
[19] E. Hu, Y. Shen *et al.*, "Lora: Low-rank adaptation of large language models," 2021.
[20] X. Han, Z. Zhang *et al.*, "Pre-trained models: past, present and future," *Ai Open*, vol. 2, pp. 225–250, 2021.
[21] J. Ni, G. Abrego *et al.*, "Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models," *arXiv preprint arXiv:2108.08877*, 2021.
[22] M. C. Hinojosa Lee, J. Braet *et al.*, "Performance metrics for multilabel emotion classification: comparing micro, macro, and weighted f1-scores," *Applied Sciences*, vol. 14, no. 21, p. 9863, 2024.
[23] M. Li, W. Ma, and Z. Chu, "User preference interaction fusion and swap attention graph neural network for recommender system," *Neural Networks*, vol. 184, p. 107116, 2025.
[24] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
[25] A. Liu, B. Feng *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.
[26] Baidu, "Ernie 3.5-8k technical introduction," https://cloud.baidu.com/doc/WENXINWORKSHOP/s/jlil56u11, 2025, accessed: 2025-07-06.