Parameter-Efficient Instruction Tuning of Large Language Models For Extreme Financial Numeral Labelling

Subhendu Khatuya[1] **Rajdeep Mukherjee**[1] **Akash Ghosh**[1] **Manjunath Hegde**[2]
Koustuv Dasgupta[2] **Niloy Ganguly**[1] **Saptarshi Ghosh**[1] **Pawan Goyal**[1]

[1]Indian Institute of Technology Kharagpur, India
[2]Language Modeling, Goldman Sachs

subha.cse143@gmail.com, rajdeep1989@iitkgp.ac.in, akashkgp@gmail.com,
manjunath.y.hegde@gs.com, Koustuv.x.Dasgupta@gs.com, niloy@cse.iitkgp.ac.in,
saptarshi@cse.iitkgp.ac.in, pawang@cse.iitkgp.ac.in

Abstract

We study the problem of automatically annotating relevant numerals (GAAP metrics) occurring in the financial documents with their corresponding XBRL tags. Different from prior works, we investigate the feasibility of solving this extreme classification problem using a generative paradigm through instruction tuning of Large Language Models (LLMs). To this end, we leverage metric metadata information to frame our target outputs while proposing a parameter efficient solution for the task using LoRA. We perform experiments on two recently released financial numeric labeling datasets. Our proposed model, **FLAN-FinXC**, achieves new state-of-the-art performances on both the datasets, outperforming several strong baselines. We explain the better scores of our proposed model by demonstrating its capability for zero-shot as well as the least frequently occurring tags. Also, even when we fail to predict the XBRL tags correctly, our generated output has substantial overlap with the ground-truth in majority of the cases.

1 Introduction

The U.S. Securities and Exchange Commission (SEC) mandates publicly traded companies to disclose periodic filings such as quarterly 10-Q & annual 10-K reports. These documents are important to finance professionals and investors who rely on SEC filings to make informed investment decisions. Each company is directed to follow the *Generally Accepted Accounting Principles* (GAAP) to report the metrics appearing in these documents and tag them using the *eXtensive Business Reporting Language* (XBRL) according to a well-defined taxonomy consisting of thousands of labels. In a recently released **FNXL** dataset (Sharma et al., 2023), such numerals are tagged from a large set of **2,794 labels**. Implementing XBRL tagging therefore requires advanced accounting skills to map financial data

to the correct XBRL concepts. This requires hiring experts to meticulously review each document and assign appropriate labels which is neither a cost-effective nor a scalable solution. Fig. 1 shows various challenges involved with the task.

Prior works and their limitations: *FiNER* (Loukas et al., 2022) formulated the task of identifying relevant numerals from financial texts and labelling them with XBRL tags as an NER problem, and used a BERT-based sequence labelling approach. However, the label-set in FiNER consists of only 139 most frequently occurring XBRL tags. For practical purposes, a much larger number of XBRL tags/labels are needed to effectively annotate the diverse types of numerals present in these documents. More recently, the authors of FNXL dataset (Sharma et al., 2023), demonstrated the poor performance of FiNER when extended to thousands of labels. They also explored an extreme classification methodology, called AttentionXML. *None of these methods, however, exploit the rich metadata information available with XBRL tags*, to improve the classification performance. Table 1 provides example of XBRL tag documentations that can help with the labelling task. Among recent methods that utilize label metadata for better results, *GalaXC* (Saini et al., 2021), is a GNN-based extreme classification approach that embeds label metadata information in its document-label graph nodes. *Label Semantics* (Ma et al., 2022) is another generic approach that leverages entity descriptions to solve the standard NER task. However, none of these methods have been applied in the financial domain. We, therefore, adapt these models to use as additional baselines for our task.

Additionally, all the methods stated above *lack the capacity to identify unseen labels during inference* as they follow a discriminative paradigm. Generative models, on the other hand, display intrinsic zero-shot capabilities if trained effectively.

Example 1

On March 7, 2019, the Company completed the Park MGM Transaction and amended the MGM-MGP Master Lease concurrent [Business combination consideration transferred 1] [Others] with which the Company paid $637.5 million, of which $605.6 million was cash and the remainder was the issuance of [Business combination consideration transferred equity interests issued and issuable] approximately 1.0 million of Operating Partnership units, to a subsidiary of MGM and, as a result, the annual rent payment to [Others] the Company increased by $50 million, prorated for the remainder of the lease year.

Example 2

As of December31, 2018, AVANGRID share capital consisted of 500,000,000 shares of common stock authorized, 309,752,140 [Common stock shares authorized] [Common stock shares issued] [Others] [Minority interest ownership percentage by parent] [Common stock par or stated value per share] shares issued and 309,005,272 shares outstanding, 81.5% of which were owned by Iberdrola, each having a par value of $0.01, [Common stock value] [Collateral securities re pledged delivered or used] for a total value of common stock capital of $3 million and additional paid in of $13,657 million.

Example 3

[Business acquisition percentage of voting interests acquired] The company originally acquired 75 % of wind te holdco on november 3, 2015 drop down assets, which were consolidated [Others] with 25 % of the net assets recorded as noncontrolling interest.

Figure 1: Demonstrating the challenges in the Extreme Financial Numeral Labelling (XFNL) task. Within a financial statement, there are scenarios where every numeral is associated with a distinct XBRL tag, such as in Example 2 (6 distinct tags). Then, there are cases where a mixture of both relevant and irrelevant numerals (tagged 'Others') coexist in the same statement, often within a very limited context, such as in Examples 1 & 3.

Tag	Documentation
common stocks shares issued	Total number of common shares of an entity that have been sold or granted to shareholders (includes common shares that were issued, repurchased and remain in the treasury). These shares represent capital invested by the firm's shareholders and owners, and may be all or only a portion of the number of shares authorized.
common stocks shares authorized	The maximum number of common shares permitted to be issued by an entity's charter and bylaws.

Table 1: Examples of (XBRL tag, documentation) pairs. We observe that the two tags differ by a single word only whereas their corresponding documentations vary significantly. We take advantage of this distinction while training our FLAN-FinXC variants. More such examples given in the Appendix A.1

In this space, LLMs have achieved impressive performances for a wide range of NLP tasks across several domains (Zhao et al., 2023), including finance (Wu et al., 2023; Yang et al., 2023).

FLAN-FinXC Framework: In this work, we show for the first time that generative models (LLMs) can achieve impressive results for the XFNL task. We systematically explore and propose FLAN-FinXC, a framework of Parameter-Efficient Instruction Tuning for Extreme Classification.

Our FLAN-FinXC framework consists of FLAN-T5 (Chung et al., 2022) models instruction-tuned with carefully-curated task-specific instructions, as shown in Fig. 2, to generate the appropriate XBRL tag documentations. We then make use of an unsupervised *Tag Matcher* module to predict the final XBRL tag for this generated documentation. We perform extensive experiments to devise a total of five different model variants as part of our proposed FLAN-FinXC framework, ranging from T5-Base to FLAN-T5-Large, and with varying training strategies. We observe that FLAN-T5-Large (instruction-tuned) achieves **9.4%** Macro-F1 gains and **3.5%** Hits@1 gains, over T5-Large (fine-

tuned) for FNXL dataset, both models being architecturally same with 780M parameters. The same trend is also observed in FiNER data. This highlights the advantages of *Instruction Tuning* similar-sized LLMs over old-fashioned fine-tuning. Given that training larger models is costly, next we experiment with parameter-efficient (PEFT) techniques, specifically Prefix Tuning (Li and Liang, 2021) and LoRA (Hu et al., 2021), to instruction-tune our FLAN-T5-Large models.

Among these, we observe that FLAN-T5-Large with LoRA outperforms the Prefix-Tuned version with **2.4%** Macro-F1 gains and **5%** gains in Hits@1, giving the best performance for both the datasets.[1]

Taking advantage of the generative paradigm, parameter-efficient instruction tuning of LLMs, as well as our financial domain-specific novelty (through the use of XBRL tag documentations to improve extreme classification performance), our best model, **FLAN-T5-Large with LoRA**, outperforms the state-of-the-art *AttentionXML* model

---

[1]Experimenting with 100 NLP tasks, (Ding et al., 2023) had also made similar observations that LoRA can sometime even outperform full fine-tuning on certain tasks.

with **39.3%** Macro-F1 gains and **17.2%** Hits@1 gains, thereby achieving new state-of-the-art results for the XFNL task. We then present several interesting analyses to investigate the reasons for its considerably better performances. We find that our model achieves impressive zero-shot Macro-F1 scores of **58.89%** for the 67 XBRL tags that were unseen during training. Even for tags that appear fewer than 5 times in the training data, our model is able to achieve 41% Macro-F1 gains and 23% Hits@1 gains compared to *AttentionXML*. Qualitatively, among the instances where we fail to predict the correct XBRL tags, in around 60% of the cases, our generated tag documentations are very close to the ground truth documentations with Jaccard Similarity scores ranging between 0.6 and 0.85. The proposed model also achieves superior performance (15.22% Macro-F1 gain and 3.3% Hits@1 gains) in case of FiNER data containing only most frequent 139 labels.

Our contributions can be summarized as follows:

- We propose FLAN-FinXC, a generative framework, consisting of a suite of instruction-tuned FLAN-T5 models, varying in model sizes and training strategies, to tackle the XFNL task, which is a practically important problem in the finance domain. Different from the prior methods for the task, our models utilize the XBRL tag documentations, instead of considering the tags as just class labels.

- In addition to comparing with the state-of-the-art methods, we adopt several prior works (Label Semantics and GalaXC) for the task, along with devising our own competitive generative baselines (T5-Base and T5-Large).

- For both FiNER as well as FNXL datasets, our best model, **FLAN-T5-Large with LoRA**, achieves huge improvements over all the baselines. Additional advantages of our best model include substantial performance over rare labels, and its zero-shot capability to tag numerals with labels unseen during training. Our dataset and codes are publicly available at `https://github.com/subhendukhatuya/FLAN-FinXC`.

2 Related Works

One of the (re)emerging applications of NLP is in the field of Named Entity Recognition (NER), particularly for identifying and categorizing various entities within text. The categorization of different
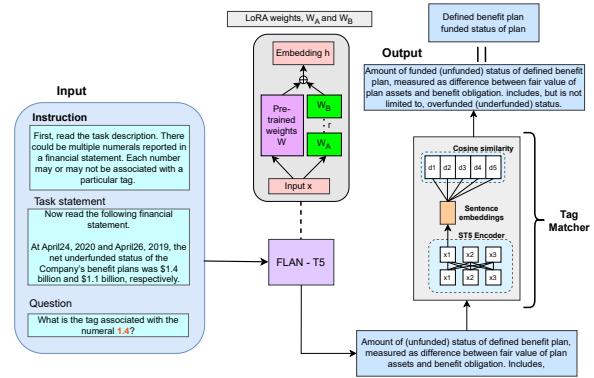

Figure 2: FLAN-FinXC Architecture. FLAN-T5 takes as input a task-specific instruction, the financial statement, and a question with a designated target numeral. FLAN-T5-generated tag documentation subsequently flows into the Tag Matcher that predicts the final tag for the given numeral.

entities into labels can also be considered as an extreme classification task (Dahiya et al., 2021). In recent years, XBRL tagging has gained a special importance in financial domain, which involves tagging of numeric values. Datasets created for such a task are the FNXL dataset (Sharma et al., 2023) and FiNER (Loukas et al., 2022) which have a very large number of entity types compared to standard NER tasks and thus present challenges to the state-of-the-art NER models. (Sharma et al., 2023) reformulated this task as an extreme classification (You et al., 2019) problem with a pipeline approach. Financial domain-specific pre-trained language models (Shah et al., 2022) have been developed, which can be repurposed on financial tasks. Among LLMs pre-trained on financial text (FinLLMs), BloombergGPT (Wu et al., 2023) is a proprietary model that suffers from limited accessibility and lack of transparency in their data collections and training protocols. FinGPT (Yang et al., 2023) is an open-source FinLLM but only suitable for financial sentiment analysis, trading, forecasting and fraud detection tasks.

3 Problem Formulation

We break the task of XFNL into two stages (with only the first stage requiring supervised training of models), as illustrated in Fig. 2, in order to take advantage of more elaborate XBRL tag documentations. We present a set of diverse annotated examples in Fig. 1. Some tag with documentation examples are shown in Table 1.

In the first stage, we formulate the problem as

a generative task using LLMs, where given a financial statement, and a question targeted towards a specific numeral occurring in the statement, the task is to accurately generate its appropriate XBRL *tag documentation* (not the tag) if the numeral is relevant and 'Other' if the numeral is irrelevant. Let $S_i = (w_i^1, ..., w_i^a, ...w_i^b, ..., w_i^n)$ be the ith statement consisting of n tokens, where w_i^a and w_i^b be two different numerals with tag_i^a and tag_i^b being their respective XBRL tag documentations. We prepend an instruction prompt IP, containing a natural language description of the task, to the statement S_i, as shown in Fig. 2. A question Q_i^a is then appended to S_i asking for the tag to be determined for a specific numeral, say w_i^a. The modified input S_i^a therefore takes the shape $IP||S_i||Q_i^a$, where $||$ is a text concatenation operation. The target answer $genTag_i^a = LLM(S_i^a)$ for the LLM (FLAN-T5 in our case) therefore becomes: tag_i^a.

In the second stage, we obtain the final XBRL tag through a separate *Tag Matcher* module, since the entire documentation may not be generated exactly. Specifically, we obtain the embedding for $genTag_i^a$ using a pre-trained state-of-the-art sentence encoder. We obtain the same for the documentations corresponding to all the available tags. The one having the highest cosine similarity with the $genTag_i^a$ embedding is declared to be the predicted XBRL tag documentation, $predTag_i^a$.

4   Methodology

Our proposed framework **FLAN-FinXC** for the XFNL task is divided into two phases, as depicted in Fig. 2: a supervised generative phase, and an unsupervised documentation-to-tag matching phase. In the first phase, given a financial statement S_i, and a question Q_i^a asking for the XBRL tag (documentation) to be determined for a numeral w_i^a appearing in the sentence S_i, we instruction tune FLAN-T5 (Chung et al., 2022; Longpre et al., 2023) with carefully-curated task-specific instructions (see Fig. 2) to generate the tag documentation $genTag_i^a$. The model is trained to condition on the modified input S_i^a, as described in the previous section, to generate the target answer one token at a time using auto-regressive decoding. Cross-entropy loss between the generated and true tokens is minimized in the process. In our case, the entire output generated by the LLM corresponds to the tag description, hence no additional parsing is required.

Our choice for FLAN-T5 is based on the observation that FLAN-T5 models are pre-trained (using instruction tuning) on more than 1.8K tasks, and hence can significantly reduce the amount of fine-tuning steps required if adopted as starting checkpoints for learning new tasks. Additionally, they achieve strong zero-shot and few-shot performances in comparison to T5 (Raffel et al., 2020), their non-instruction-tuned counterpart.

Note that we train FLAN-T5 models to generate the *XBRL tag documentations* instead of the tag themselves (which is the final target) since the more elaborate documentations allow for a better distinction than their corresponding tags (see Table 1), thereby aiding in the extremely challenging XFNL task (as demonstrated in Fig. 1). Generating the lengthy tag documentations exactly is however difficult, and hence the generated documentation $genTag_i^a$ may not exactly match the ground truth. In the second phase of our proposed framework, we therefore leverage a pre-trained state-of-the-art sentence encoder, as the backbone of our *Tag Matcher* module, as depicted in Fig. 2. More specifically, we use Sentence-T5-XXL (Ni et al., 2021), pre-trained using contrastive loss, to generate 786-dim embeddings for $genTag_i^a$ as well as for each of the ground truth tag documentations. The ground truth tag documentation with the highest cosine similarity (between embeddings) with $genTag_i^a$ is considered to be the predicted tag documentation $predTag_i^a$. Since there exists a 1:1 mapping between XBRL tags and their corresponding documentations, the final predicted tag can be easily obtained from $predTag_i^a$.

In order to investigate the suitability of a generative paradigm for solving the task, we perform a systematic evaluation of several model variants, varying both model sizes and training strategies.
Non-FLAN Model Variants: First, we compare the fine-tuned performances of T5-Base (220M parameters) with T5-Large (780M parameters). Note that for fine-tuning, given a financial statement S_i, the modified input now becomes $S_i^a = S_i||Q_i^a$, where Q_i^a refers to the question targeted towards the numeral w_i^a appearing in S_i. The target (tag documentation) however remains the same.
FLAN-FinXC Model Variants: Next, we compare the instruction-tuned performance of FLAN-T5-Large with fine-tuned T5-Large, both models with identical architectures. Next, we instruction tune FLAN-T5-Large with PEFT techniques, namely Prefix Tuning and LoRA, respectively re-

quiring only 0.13% and 0.08% of model parameters to be updated. Finally, we instruction-tune FLAN-T5-Large with LoRA to achieve our best results.

5 Baselines

We compare the performance of our proposed models with the following baselines:
Baselines already applied for XBRL tagging: **(1) FiNER:** (Loukas et al., 2022) solved the task as a NER task, but only for top 139 frequent XBRL tags. We adapt this method for our use case (large number of labels). **(2) AttentionXML Pipeline:** (Sharma et al., 2023) tackled this problem using a BERT-based sequence-to-sequence tagger.
New baselines that we adopt for XBRL tagging: Additionally, we adopt two diverse methods for the task for XBRL tagging, which can utilize the label semantics of XBRL tags. **(3) GalaXC:** (Saini et al., 2021) applies collaborative learning over document-label graphs that allows additional information like label metadata to be incorporated. **(4) Label Semantics** (Ma et al., 2022) leverages the entity description to solve a standard NER task. Lastly, with the emergence of ChatGPT(OpenAI, 2022), we were curious to check ChatGPT's performance for this task for 500 random samples using *gpt-3.5-turbo* [2] API.

6 Dataset & Evaluation Metrics

Dataset: We perform our experiments on the recently released FNXL dataset (Sharma et al., 2023) containing 10-K documents[3] for 2,339 companies and FiNER (Loukas et al., 2022) dataset having most frequent (at least 1000 appearances) 139 labels. FNXL dataset contains a total of 79,088 sentences containing 142,922 annotated numerals, with a tag set of *2,794 distinct tags* that follows a heavy tail distribution. Further, the test set contains *67 XBRL tags that are not part of the train or validation sets*. Fig. 1 shows a few annotated examples from this dataset. Each tag is associated with a pre-defined textual description known as its 'tag documentation' (see Table 1). The tag documentations have an average length of 28 words and a maximum length of 273 words.
Metrics: To ensure a comprehensive evaluation of all models, we employ the following metrics: 1) Macro-Precision, 2) Macro-Recall, 3) Macro-F1,

---

[2] https://platform.openai.com/docs/models/gpt-3-5

[3] https://tinyurl.com/t43mwd5m

4) Hits@1. In financial numeral labelling, where equal importance is given to all tags, macro based metrics (precision, recall, F1) are a good choice as it treats all classes equally, irrespective of their frequency. We report *Hits@1* metric to showcase the usability of this system from business perspective, recommending the *top tag* to domain experts.

7 Experimental Setup

For all our model variants (experiments performed on Tesla V100 32GB GPUs), we obtain the pre-trained checkpoints from the *Huggingface* Library[4]. For instruction tuning FLAN-T5-Large with Prefix Tuning, the *prefix length* was set to 20. For training the models with LoRA, the *rank* for the trainable decomposition matrices was set to 2. FLAN-T5-Large models were instruction-tuned for 10 epochs with a learning rate (lr) of $1e-4$ (training time: 1hr 22 minutes/epoch, inference time: 2 minutes/sample); with an lr of $1e-2$ with Prefix Tuning (training time: 29 minutes/epoch, inference time: 2 minutes/sample); and with an lr of $5e-4$ with LoRA (training time: 56 minutes/epoch, inference time: 2 minutes/sample). These hyperparameters were selected based on the best Macro-F1 results on the validation set.

Note that the tag information is *not* included in the input prompt. Rather, it is generated during the decoding phase based on the question asked. The input prompt containing the question gets encoded and it is much smaller than the maximum input context length. Hence, there is no truncation during encoding. The average tag information length is 28 tokens and we set the length of output to be generated to 30 tokens. Hence, in a few cases, truncation may occur during the decoding phase. However, the crucial tag information is typically present in the initial tokens, and our tag matcher ensures correct labels by comparing generated tags with the descriptions for all available tags in the dataset. Thus any potential truncation in a few cases does not pose any risk to our model's performance.
Instruction Selection: While experimenting, we gradually improved our instructions. First, we tried out a prompt without any task-specific instruction, thereby consisting of the input financial statement and the question for a given numeral. Then, we included a simple task description and repeated our experiments. Finally, we made the task description/instruction more elaborate which led us to ob-

---

[4] https://huggingface.co/

tain the best results. On the target side, we tried out two variants, generating labels vs. generating the label descriptions. The latter resulted in significant performance improvement (refer Table 4). This iterative process of instruction selection was guided by the performance on the validation set.

8 Main Results

We report the results of our proposed model variants and various baselines in Table 2 for both FNXL and FiNER dataset. Among the baselines, FiNER does not perform well for large number of entity labels. On the other hand, while the 'Label Semantics' method leverages tag words within an NER framework, its scalability is constrained when dealing with a large number of entity labels, ultimately leading to poor performance. The AttentionXML pipeline performs better than all other baselines. GalaXC and 'Label Semantics' have very similar performance, with GalaXC giving slight edge. ChatGPT's performance for this complex task is not satisfactory.

We also demonstrate the advantage of training bigger models for extreme financial numeral labelling for FNXL dataset, by comparing the results of T5-Large (780M parameters, Macro-F1 60.11) with T5-Base (220M parameters, Macro-F1 54.21). Note that these baselines that we devise for the task, already outperform the existing state-of-the-art.

Next, we turn to different variations of our FLAN-FinXC framework (listed in the lower part of Table 2). First, we demonstrate the advantage of instruction tuning over fine-tuning by the fact that FLAN-T5-Large (Macro-F1 65.77) substantially outperforms T5-Large (Macro-F1 60.11) for FNXL dataset. Given that training larger models is costly, next we experiment with parameter-efficient (PEFT) techniques, specifically Prefix Tuning (Li and Liang, 2021) and LoRA (Hu et al., 2021), to instruction tune our FLAN-T5-Large models. Both techniques require only a small fraction of model parameters to be fine-tuned (0.13% with Prefix Tuning, and 0.08% with LoRA), thereby greatly reducing the computation cost and training times. Instruction tuning FLAN-T5-large with LoRA (Macro-F1 66.23) outperforms Prefix-Tuned version of the same model (Macro-F1 64.65) and has a slight edge over the model variant instruction-tuned without PEFT (Macro-F1 65.77). Our best results are obtained by instruction tuning FLAN-T5-Large with LoRA (only 0.08% of 780M param-

eters to be finetuned). This model outperforms the state-of-the-art *AttentionXML* model with **39.3%** Macro-F1 gains and **17.2%** Hits@1 gains, thereby achieving new state-of-the-art results for FNXL dataset. For FiNER dataset, we achieve 15.22% Macro-F1 gain and 3.3% Hits@1 gains.

9 Analysis

We now present different analyses of our best model (FLAN-T5-Large with LoRA), and the closest baseline (AttentionXML Pipeline). The following analysis is specific to the FNXL dataset, chosen for its more extensive range of labels compared to other datasets, emphasizing its practical relevance.

9.1 Performance on least frequent labels

Accurate tagging of infrequent labels in XBRL is crucial for reliable financial reporting, yet it presents challenges stemming from imbalanced data, scarce training instances, and sparse data distribution. We observe our FLAN-FinXC performs substantially better than AttentionXML for the rare labels. To demonstrate this, we group the tags into various buckets based on their frequency of occurrence in the training set.
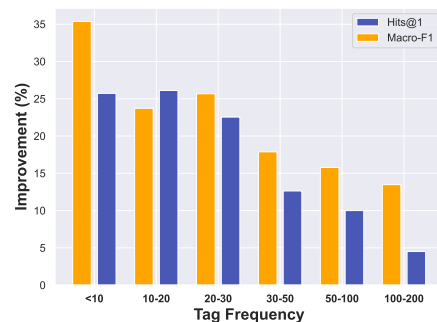

Figure 3: Relative improvement in performance achieved by FLAN-T5-Large with LoRA over AttentionXML Pipeline, for the least frequent labels under various frequency buckets

Fig. 3 shows the percentage improvement in Hits@1 and Macro-F1 that is achieved by FLAN-FinXC over AttentionXML Pipeline for the tags in every bucket. We see that our model can effectively identify and tag rare financial concepts. Notably, even for tags that appear fewer than 10 times in the training data, our model is able to achieve 35.3% improvement in Macro-F1 score and 25% in Hits@1 over the closest baseline.

Model	Dataset							
	FNXL (Sharma et al., 2023)				FiNER (Loukas et al., 2022)			
	M-P	M-R	M-F1	Hits@1	M-P	M-R	M-F1	Hits@1
FiNER (bert-base)	49.17	<u>49.71</u>	47.13	75.34	72.60	81.10	76.61	81.50
FiNER (sec-base)	47.76	48.87	46.20	74.67	81.11	<u>83.20</u>	82.14	82.30
Label Semantics	46.35	45.12	45.72	71.25	71.50	80.15	75.57	80.25
GalaXC	46.91	44.81	45.81	72.97	72.20	80.95	76.32	81.10
AttentionXML Pipeline	<u>50.69</u>	48.51	<u>47.54</u>	<u>76.76</u>	<u>82.15</u>	82.30	<u>82.22</u>	<u>83.25</u>
ChatGPT (500 samples)	11.13	7.68	9.08	19.6	20.12	15.67	17.61	22.35
T5-Base	59.94	49.48	54.21	79.21	86.92	84.35	85.61	83.45
T5-Large	61.87	58.46	60.11	83.26	88.12	85.10	86.58	84.12
FLAN-T5-Large	**66.21**	65.34	65.77	86.21	**92.10**	96.35	94.17	85.89
FLAN-T5-Large with Prefix-Tuning	65.10	64.21	64.65	85.69	90.18	94.35	92.21	85.12
FLAN-T5-Large with LORA	65.14	**67.36**	**66.23**	**89.98**	91.84	**97.85**	**94.74**	**86.03**

Table 2: Performance evaluation based on Macro & Hits@1 metrics for FNXL dataset and FINER dataset. The best performance is highlighted in bold, and the strongest baseline result is underlined.

9.2 Zero-Shot Capability

One of the key strengths of our proposed model is its zero-shot capability, i.e., its ability to generate tags/labels for which it has *not* been explicitly trained. While SOTA discriminative models in this domain often require specific fine-tuning or retraining to handle new tags, our generative model transcends these limitations.

Tag	F1-score
foreign currency transaction gain before tax	0.85
commercial paper at carrying value	0.85
accounts payable other current	0.80
recognition of deferred revenue	0.76
available for sale debt securities gross unrealized gain	0.66

Table 3: Zero-Shot Performance of FLAN-T5-Large with LoRA, for a few XBRL tags absent in the training set of FNXL dataset.

Recall that the test set included 67 new labels that were not present in the train set. We observe that our best model achieves a Macro-F1 of **58.89** over these 67 unseen labels, which is a commendable performance. Table 3 shows the performance (F1-score) of our best model for few such tags.

9.3 Ablation Study

We now try out various ablations over our best model to understand the significance of different modules. First, as ablations of our *Tag Matcher* module, we replace Sentence-T5-XXL (used in our best model) with Sentence-BERT (Reimers and

Model	Macro-F1	Hits@1
FLAN-T5-Large with LoRA	66.23	89.98
w/ S-BERT-L12 as Tag Matcher	63.11	88.13
w/ S-BERT-L6 as Tag Matcher	62.87	87.72
w/o instruction prompt	56.46	76.55
w/o tag metadata	53.12	73.14

Table 4: Results for our ablation studies of FNXL dataset

Gurevych, 2019) to generate the embeddings for the XBRL tag documentations. We experiment with two versions of Sentence-BERT, one with 6 encoder layers and other with 12 encoder layers, and report our findings in Table 4. We observe a drop in performance (compared to our best model FLAN-FinXC with LoRA) in both cases, thereby showing the better efficacy of Sentence-T5-XXL over Sentence-BERT as sentence encoder.

Next, we instruction-tune FLAN-T5-Large (with LoRA) *without the instruction prompt* containing specific instructions for the XFNL task. Given a financial statement S_i, the modified input now becomes $S_i^a = S_i || Q_i^a$, where Q_i^a refers to the question targeted towards the numeral w_i^a appearing in S_i, and $||$ is a text concatenation operation. From the fourth row in Table 4, we observe a huge drop in performance (e.g., Macro-F1 drops from 66.23 to 56.46), thereby demonstrating the importance of aligning FLAN-T5 fine-tuning with task-specific instructions to tackle the challenging XFNL task.

Next, we experiment with setting the XBRL tags, instead of their documentations, as the target for

FLAN-T5-Large. In other words, we no longer use the tag metadata. Accordingly, the *Tag Matcher* module now compares the embeddings of the generated and ground truth tags (and not documentations). From the last row in Table 4, we again observe a drop in scores with 19.8% ↓ in Macro-F1 and 18.7% ↓ in Hits@1 compared to our best results. This confirms our hypothesis that the more elaborate tag documentations allow for a clearer distinction than the corresponding tags (possibly differing only in few words) while tackling the challenging extreme classification task.

Sentence	Ground truth tag	AttentionXML Pipeline prediction	FLAN-T5-Large with LoRA
As of December 31, 2019 and 2018, we had a cumulative translation loss, net of tax of $ 617 million and $ 466 million, respectively.	accumulated other comprehensive income loss foreign currency translation adjustment net of tax	accumulated other comprehensive income loss defined benefit pension and other postretirement plans net of tax ✗	accumulated other comprehensive income loss foreign currency translation adjustment net of tax ✓
We also have $4.54 billion of non - cancelable contractual commitments as of December 31, 2019 related to network infrastructure	contractual obligation	unrecorded unconditional purchase obligation balance sheet amount ✗	contractual obligation ✓
At April 24, 2020, plan participants had approximately $ 14 million withheld to purchase the company's ordinary shares	others	share based compensation arrangement by share based payment award number of shares available for grant ✗	others ✓

Table 5: Challenging examples where AttentionXML predicts a wrong tag, while FLAN-T5-Large with LoRA predicts correctly.

Qualitative Analysis of predicted labels: Table 5 shows the predictions made by our best model and the closest baseline (AttentionXML) for a few challenging instances. Our proposed method classifies relevant numerals (first two examples) correctly and is able to find irrelevant numerals (the last example) and correctly tag that as *'Others'*, whereas the baseline struggles for each cases.

Error Analysis: To characterize the errors committed by our model, we ask – *when a model generates a wrong tag, how similar is the generated tag with the ground truth tag?* We quantify the similarity between a generated tag and the ground truth (GT) tag by the Jaccard similarity between the tag documentation words. Fig. 4 compares the errors by our best model and errors by the closest baseline. For a majority (60%) of errors by our best model, the

Sentence: Our Board of Directors declared quarterly dividends per share of $1.45, $1.32 and $1.15, which were paid in each of the four quarters of 2019, 2018, and 2017.
GT Tag Doc: aggregate dividends paid during the period for each share of common stock outstanding.
GT Tag: common stock dividends per share cash paid
FLAN-T5-Large with LoRA (Generated, No Tag Matcher): Aggregate dividends declared during the period for each share of common stock outstanding.
Predicted Tag: common stock dividends per share declared

Sentence: In fiscal year 2016, Bard paid the Company $121 million towards the settlement of 11,000 of these claims.
GT Tag Doc: amount awarded from other party in judgment or settlement of litigation.
GT Tag: litigation settlement amount awarded from other party
FLAN-T5-Large with LoRA (Generated, No Tag Matcher): Amount awarded to other party in judgment or settlement of litigation
Predicted Tag: litigation settlement amount awarded to other party

Table 6: Examples to show that subtle differences between ground truth tag (GT) and the predicted tag. The few-word differences between Ground Truth (green) and predicted text (red) are highlighted. More such examples are in the Appendix A.3

wrongly predicted tag is very similar to the ground truth tag (Jaccard similarity between tags ≥ 0.6). Whereas, most of the tags wrongly predicted by AttentionXML are very different from the ground truth tags (Jaccard similarity between tags ≤ 0.4).
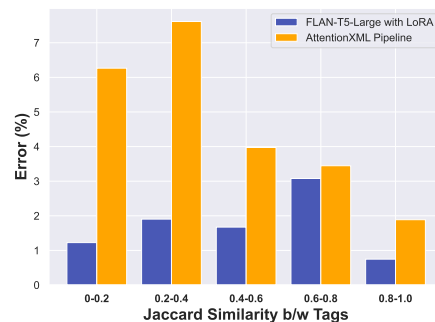

Figure 4: Comparing errors by best proposed model and those by the closest baseline. Even when our model generates incorrect tags, most of them are semantically very similar to the ground truth tags. But AttentionXML often generates completely unrelated tags.

Finally, Table 6 demonstrates instances where the predicted tag documentation closely resembled the ground truth (GT) tag documentation, but even minor variations led to a wrong final tag prediction. As illustrated in Table 6, subtle differences between *'from other party'* (in the GT tag, highlighted in green) and *'to other party'* (in the prediction, highlighted in red), or between *'dividends paid'* (in the

GT tag, highlighted in green) and *'dividends declared'* (in the prediction, highlighted in red) can lead to a wrong final tag prediction. To address these complexities, in future we plan to incorporate external financial knowledge.

9.4 Comparison with ChatGPT

We used ChatGPT (GPT-3.5 turbo) under a few-shot setting. For a fair comparison, we used the same prompt (shown in Fig. 2) that is used to instruction tune our proposed models. The ChatGPT prompt structure was: <*Instruction, Task statement, Question for a specific numeral*>. In order to guide ChatGPT what to generate, we augment the prompt with 5 exemplars of (input, desired output) pairs, thereby making it a 5-shot setting. Based on experimental observations, we set "desired outputs" to be the labels instead of label descriptions. Hence, in our final setting, no label descriptions were fed. On the same 500 test set samples used to report ChatGPT's performance (macro-f1 of 9.08) in Table 2, we obtain macro-f1 score of **58.25** of our best model when trained to generate labels (note that Table 4 last row reports scores on the full test set). We believe that ChatGPT's lower performance is owing to two main factors: 1) It is not fine-tuned; rather, we use it in a 5-shot (few-shot) setting. 2) ChatGPT was not specifically pretrained on financial data and the challenging nature of the FNXL dataset further makes it difficult. Further analysis can be found in Appendix A.2.

9.5 Experimental comparison among models

As stated previously, given the unavailability of generative baselines for the task, we had to first train and compare a suite of generative baselines as reported in Table 2. This initial step allowed us to establish a fair comparison among models of similar sizes before showcasing the incremental performance gains with larger models. Table 7 presents a comparison of the number of trainable parameters in the different baselines, including the ones (generative) we trained, and our best-performing model. Also, a comparison of training times per epoch of the different models is shown in Table 8. T5-Base (220M), the weakest generative baseline trained by us, is comparable with Label Semantics, the largest existing non-generative baseline (218M). From Table 2, T5-Base already outperforms Label Semantics with **18.56%** Macro-F1 and **11.17%** Hits@1 gains. T5-Base also outperforms AttentionXML (112M) substantially as reported above.

Also, while FLAN-T5-Large (780M) is larger than AttentionXML, our best results are obtained when we instruction-tune FLAN-T5-Large with LoRA, thereby drastically reducing the number of trainable parameters (0.59M), substantially lesser than AttentionXML (112M). Yet, we significantly outperform AttentionXML with **39.3%** Macro-F1 and **17.2%** hits@1 gains.

Model	Trainable parameters
FINER	109M
Label Semantics	219M
GalaXC	41M
AttentionXML	112M
T5-Base	220M
T5-Large	739M
Flan-T5-Large	780M
Flan-T5-Large w/ LoRA	**0.59M**

Table 7: Comparison of number of trainable parameters between the baselines and FLAN-T5-Large with LoRA

Model	Training time per epoch
FINER	12 hours
Label Semantics	9 hours
GalaXC	1.2 hours
AttentionXML	4 hours
Flan-T5-Large w/ LoRA	**56 minutes**

Table 8: Comparison of training time per epoch between the baselines and FLAN-T5-Large with LoRA

10 Conclusion

This work proposes a generative approach to solve the financial numeric labelling task. We propose a novel FLAN-FinXC framework, that makes use of parameter-efficient instruction tuning of LLMs for this extreme labelling task. While comparing with the state-of-the-art models and various competitive baselines that we devise for the task, we find that our best model, Flan-T5-Large with LoRA, achieves huge improvements, providing a Macro-F1 of 66.23% as compared to the previously reported best numbers of 47.54% for FNXL dataset. For adapting to other tasks, our approach requires only minor task-specific modifications to the input prompt, and the target output. Moreover, our approach offers flexibility in using either the class labels or label descriptions, depending on the available data. As potential future directions, we believe the scope to include more financial knowledge and integrate a human-AI feedback loop would be the way forward to improve performance of this challenging task.

11 Limitations

In this work, we have not integrated external financial knowledge to address the subtle differences between tags as identified in our error analysis. Also, we have observed that labeling numerals solely based on sentence-level text (as done in this work) can be challenging, since the context depends on the surrounding paragraph, associated tables, and other elements which are not used in this work. Incorporating such elements as well as external financial domain knowledge into a financial numeral labeling model would be interesting future works.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Kunal Dahiya, Deepak Saini, Anshul Mittal, Ankush Shaw, Kushal Dave, Akshay Soni, Himanshu Jain, Sumeet Agarwal, and Manik Varma. 2021. Deepxml: A deep extreme multi-label learning framework applied to short text documents. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 31–39.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. 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)*, pages 4419–4431.

Jie Ma, Miguel Ballesteros, Srikanth Doss, Rishita Anubhai, Sunil Mallya, Yaser Al-Onaizan, and Dan Roth. 2022. Label semantics for few shot named entity recognition. *arXiv preprint arXiv:2203.08985*.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.

TB OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. openai.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Deepak Saini, Arnav Kumar Jain, Kushal Dave, Jian Jiao, Amit Singh, Ruofei Zhang, and Manik Varma. 2021. Galaxc: Graph neural networks with labelwise attention for extreme classification. In *Proceedings of the Web Conference 2021*, pages 3733–3744.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*.

Soumya Sharma, Subhendu Khatuya, Manjunath Hegde, Afreen Shaikh, Koustuv Dasgupta, Pawan Goyal, and Niloy Ganguly. 2023. Financial numeric extreme labelling: A dataset and benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3550–3561.

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

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models.

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 32.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu,

Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

A Appendix

In this section, we provide some supplementary materials that enhance the content presented in the main paper.

A.1 XBRL Tag Documentation

Each XBRL tag is associated with a pre-defined textual description known as its 'tag documentation'. Table 9 shows some more examples of XBRL tags and their documentations. Note that, while certain tag-pairs may exhibit subtle distinctions, their accompanying documentations vary significantly. Domain experts leverage the information provided in the documents during manual annotation. Our proposed model is also designed to utilize these documentations.

A.2 Comparison with ChatGPT

With the emergence of ChatGPT, we were curious to check its performance on this task using the same instruction prompt for a few samples. As can be seen from the examples in Table 10, Chat-GPT struggles to correctly identify the appropriate tags in the majority of cases, and in some cases, it misunderstands the context entirely. This indicates that ChatGPT has still very limited performance in financial numerical tagging.

A.3 Error Analysis

Table 11 demonstrates instances where the generated tag documentation (by our best model) closely resembled the ground truth (GT) tag documentation, but even minor variations led to a wrong final tag prediction. In the financial domain, subtle word changes can drastically alter the context. As illustrated in Table 6, subtle differences between *'from other party'* (in the GT tag, highlighted in green) and *'to other party'* (in the prediction, highlighted in red), or between *'dividends paid'* (in the GT tag, highlighted in green) vs *'dividends declared'* (in the prediction, highlighted in red), or between *'cash outflow'* and *'cash inflow'*, *'loss'* vs *'damages'* can lead to a wrong final tag prediction.

Tag	Documentation
common stocks shares issued	Total number of common shares of an entity that have been sold or granted to shareholders (includes common shares that were issued, repurchased and remain in the treasury). These shares represent capital invested by the firm's shareholders and owners, and may be all or only a portion of the number of shares authorized. Shares issued include shares outstanding and shares held in the treasury.
common stock shares authorized	The maximum number of common shares permitted to be issued by an entity's charter and bylaws.
tax credit carry forward amount	The amount of the tax credit carryforward, before tax effects, available to reduce future taxable income under enacted tax laws.
tax credit carry forward valuation allowance	Amount of valuation allowance pertaining to the deferred tax asset representing potential future taxable deductions from tax credit carryforwards for which it is more likely than not that a tax benefit will not be realized.
due to affiliate noncurrent	Amount of receivables owed to an entity that is affiliated with the reporting entity by means of direct or indirect ownership, which are usually due after one year (or one business cycle, if longer).
due to affiliate current and non current	Amount of payable due to an entity that is affiliated with the reporting entity by means of direct or indirect ownership.
notes payable related parties classified current	The amount for notes payable (written promise to pay), due to related parties. Used to reflect the current portion of the liabilities (due within one year or within the normal operating cycle if longer).
notes payable related parties current and non current	The amount for notes payable (written promise to pay), due to related parties.

Table 9: Examples of XBRL tag documentations. While some tag-pairs exhibit very subtle distinctions, their accompanying documentations vary significantly. Our model takes advantage of the distinctions between the tag documentations.

Instruction: First, read the task description. There could be multiple numerals reported ....
Sentence: Now read the following financial statement.
At April24, 2020 the estimated fair value was $**27.1** billion compared to a principal value of$24.5 billion.
Question: What is the tag associated with the numeral **27.1**?
GT Tag: long term debt fair value
ChatGPT: estimated fair value
FLAN-FinXC: long term debt fair value

Sentence: Ordinary shares - par value $0.0001, **2.6** billion shares authorized, 1,345,400,671 and 1,341,074,724 shares issued and outstanding, respectively
Question: What is the tag associated with the numeral **2.6**?
GT Tag: common stock shares authorized
ChatGPT tag: ordinary shares authorized
FLAN-FinXC: common stock shares authorized

Sentence: Share Capital Medtronic plc is authorized to issue 2.6 billion Ordinary Shares, $0.0001 par value; 40 thousand Euro Deferred Shares, €1.00 par value; 127.5 million Preferred Shares, $0.20 par value; and **500** thousand A Preferred Shares, $1.00 par value.
Question: What is the tag associated with the numeral **500**?
GT Tag: preferred stock shares authorized
ChatGPT: a preferred shares authorized
FLAN-FinXC: preferred stock shares authorized

Instruction: ...... **Sentence:** At April26, 2019, $764 million of rebates were classified as other accrued expenses and $**432** million of rebates were classified as a reduction of accounts receivable in the consolidated balance sheets.
Question: What is the tag associated with the numeral **432**?
GT Tag: contract with customer refund liability current
ChatGPT: rebates reduction of accounts receivable
FLAN-FinXC: contract with customer refund liability

Sentence: As of December 31, 2018, we expect to receive total future rental income of $**203** million related to noncancelable subleases for abandoned facilities.
Question: What is the tag associated with the numeral **203**?
GT Tag: operating leases future minimum payments receivable
ChatGPT tag: future rental income from noncancelable subleases
FLAN-FinXC: operating leases rent expense sublease rentals 1

Table 10: Comparison between ChatGPT and our best FLAN-FinXC model variant's prediction

Instruction: First, read the task description. There could be multiple numerals reported in a financial statement......**Sentence:** On January 15, 2020, the parties agreed to a settlement in principle to resolve the lawsuit, which will require a payment of $**550** million by us and is subject to approval by the court.
Question: What is the tag associated with the numeral **550**?
GT Tag Doc: amount of loss contingency liability. **GT Tag:** loss contingency accrual at carrying value
Flan-Large Generated: Amount awarded damages contingency liability.
Flan-Large +Tag Matcher: amount of damages awarded to the plaintiff in the legal matter.
Pred Tag-Words: loss contingency damages awarded value

Sentence: In fiscal year 2016, Bard paid the Company $**121** million towards the settlement of 11,000 of these claims. **Question:** What is the tag associated with the numeral **121**?
GT Tag Doc: amount awarded from other party in judgment or settlement of litigation. **GT Tag:** litigation settlement amount awarded from other party
Flan-Large Generated: Amount awarded to other party in judgment or settlement of litigation
Flan-Large +Tag Matcher: amount awarded to other party in judgment or settlement of litigation.
Pred Tag-Words: litigation settlement amount awarded to other party

---

Sentence: During the year ended December 31, 2017, we issued and repaid an aggregate of $**12.3** billion of commercial paper and had a maximum outstanding balance of $1.5 billion under our commercial paper program. **Question:** What is the tag associated with the numeral **12.3**?
GT Tag Doc: the cash outflow due to repaying amounts borrowed by issuing commercial paper.**GT Tag:** repayments of commercial paper
Flan-Large Generated: The cash inflow during to theing short borrowed under issuing commercial paper.
Flan-Large +Tag Matcher: the cash inflow from borrowing by issuing commercial paper.
Pred Tag-Words: proceeds from issuance of commercial paper

Sentence: Our Board of Directors declared quarterly dividends per share of $**1.45**, $1.32 and $1.15, which were paid in each of the four quarters of 2019, 2018, and 2017, respectively. **Question:** What is the tag associated with the numeral **1.45**?
GT Tag Doc: aggregate dividends paid during the period for each share of common stock outstanding.
GT Tag: common stock dividends per share cash paid
Flan-Large Generated: Aggregate dividends declared during the period for each share of common stock outstanding.
Flan-Large +Tag Matcher: aggregate dividends declared during the period for each share of common stock outstanding.
Pred Tag-Words: common stock dividends per share declared

Instruction: .....**Sentence:** As of December 31, 2020, Duong met the held-for-sale criteria and loan receivable balance of $1.3 billion, net of CECL reserve of $**32** million was reclassified **Question:** What is the tag associated with the numeral **32**?
GT Tag Doc: amount of allowance for credit loss on accounts receivable. **GT Tag Words:** allowance for doubtful accounts receivable
Flan-Large Generated: othersmount of allowance for credit loss on financing receivable, A
Flan-Large +Tag Matcher: amount of allowance for credit loss on financing receivable, classified as noncurrent.
Pred Tag-Words: allowance for notes and loans receivable noncurrent

Table 11: Error cases. Examples to show the subtle difference between ground truth (shown in green color) and generated tag docs (shown in red color) and predicted tag.