

# Exploring LLMs for Financial Numeric Labelling and Summarization

Subhendu Khatuya Computer Science and Engineering Department, IIT Kharagpur  
subha.cse143@gmail.com

## ABSTRACT

We tackle the challenge of automatically annotating GAAP metrics in financial documents with XBRL tags using a generative approach through instruction tuning of Large Language Models (LLMs). By leveraging metric metadata and employing a parameter-efficient solution via LoRA, our model, FLAN-FinXC, achieves state-of-the-art performance on two financial numeric labeling datasets, outperforming strong baselines. Additionally, we explore the summarization of long Earning Call Transcripts (ECTs) using the ECT-Sum dataset. Our model, FLAN-FinBPS, combines an unsupervised question-based extractive module with an instruction-tuned abstractive module, achieving a 14.88% average ROUGE score gain over the strongest baseline and producing factually consistent bullet point summaries.

## KEYWORDS

LLMs, Instruction-Tuning, Summarization, Unsupervised

## 1 INTRODUCTION

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 entities into labels can also be considered as an extreme classification task. 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 and FiNER 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. 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[3], 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. First, we study the problem of automatically annotating relevant numerals occurring in the financial documents with their corresponding XBRL tags and propose a novel generative approach.

In finance, another important task is to quickly summarize the key facts of long Earnings Calls to take any financial decisions. Earnings Calls, typically a teleconference or a webcast, are hosted by publicly

traded companies to discuss important aspects of their quarterly (10-Q), or annual (10-K) earnings reports, along with current trends and future goals that help financial analysts and investors to review their price targets and trade decisions. The corresponding call transcripts (called Earnings Call Transcripts, abbreviated as ECTs) are typically in the form of long unstructured documents consisting of thousands of words. Hence, it requires a great deal of time and effort, even on the part of trained analysts, to quickly summarize the key facts covered in these transcripts. While automatic summarization techniques have made significant advancements, their primary focus has been on summarizing short newswire articles or documents that have clear structural patterns like scientific articles or government reports. Unfortunately, there hasn't been much exploration into developing efficient methods for summarizing financial documents, which often contain complex facts and figures. Here, we study the problem of bullet point summarization of long earning call transcripts using recently released ECTSum dataset and propose a novel method. ECTSum is a challenging dataset for automatic bullet point summarization primarily due to the following factors: (i) it has a very high document-summary compression ratio of 103.67; (ii) ECTs, with an average document length of approximately 2.9K words, are free form long documents as opposed to other existing long document summarization datasets with a fixed document layout; (iii) average token count of ECTs are much more than the maximum token limit of common LLMs such as BERT, T5 or BART.

## 2 RESEARCH QUESTIONS AND CONTRIBUTIONS

In my PhD, in the last three years, I have addressed the following Research Questions and exploring further.

**RQ1: Exploring Generative Paradigm for Extreme Financial Numeral Labelling:** In the first stage, we formulate the problem as a generative task using LLMs. Let  $S_i = (w_i^1, \dots, w_i^a, \dots, w_i^b, \dots, w_i^n)$  be the  $i^{th}$  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  $S_i$ . 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 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. Note that, after extensive experimentation with different prompts, we have determined the optimal task-specific prompt for this.

**Contribution: RQ1 - Solution Overview [Khatuya et al.,2024]:** Available state-of-the-art models for XFNL task *lack the capacity*

to identify unseen labels during inference as they follow a discriminative paradigm. 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, 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 most of the cases.

**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[1], a framework<sup>1</sup> of Parameter-Efficient Instruction Tuning for Extreme Classification. Our framework consists of FLAN-T5 models instruction-tuned with carefully-curated task-specific instructions, to generate the appropriate XBRL tag documentations. We then use 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 framework, ranging from T5-Base to FLAN-T5-Large, and with varying training strategies.

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 high Jaccard Similarity scores.

**RQ2: Bullet Point Summarization of Long Earnings Call Transcripts** Here, I have worked on bullet point summarization of ECT reports on a recent new benchmark dataset ECTSum<sup>2</sup> containing 2,425 ECT report-summary pairs. The target summaries in their dataset are extremely concise bullet-point style summaries. Given a long ECT, our goal is to propose an efficient and novel method to create a concise factually consistent bullet-point summary.

**Contribution: RQ2 - Solution Overview [Khatuya et al.,2024]:** In this work, we propose a novel two-stage generative framework **FLAN-FinBPS**[2], that uses a combination of unsupervised and supervised methods to produce abstractive bullet point summaries of ECT documents. As opposed to the previous state-of-the-art method which used supervised fine-tuned approaches in both the stages of the model, we use an unsupervised question-based context generator module to produce the extractive summary in the first stage, thereby cutting down on the training time of our model. The second stage of our framework utilises a supervised parameter-efficient instruction-tuned module to generate the abstractive bullet point summaries by

using the extractive summary as the context. Typically, each bullet point summary highlights 3-4 crucial financial aspects of the input ECT document, such as revenue, income, earnings per share, sales, profit, equity, etc. This observation motivated us to initially identify the significant *topics* present in a given ECT. We first generate a list of *questions* for each ground truth bullet point summary in the train set using an unsupervised pre-trained T5 model. For each sentence in the ground truth summary, we generate a corresponding question.

Our model<sup>3</sup> outperforms the strongest baseline, achieving a notable **14.88%** increase in average ROUGE score and a **16.36%** rise in BERTScore, signifying a major enhancement in content quality. It also generates more precise numerical values, showcasing a **2.51%** gain in Num-Prec, and produces more factually consistent summaries, demonstrating a **2.70%** gain in SummaC<sub>CONV</sub> compared to the previous strongest state-of-the-art method.

**RQ3: Instruction-Guided Numerical Reasoning of Financial Data**

Since their inception, large language models (LLMs) have taken steady strides in terms of what they can do. As they get more and more integrated in our everyday lives, there are areas where LLMs still lag behind; one such area being numerical reasoning. It is therefore a focus of active research to increase the reasoning capabilities of LLMs. Prompt based techniques such as chain of thought prompting, tree of thought prompting, program of thought prompting have been proposed to this end. These techniques aim to tell the LLM how to think or reason by stating a series of intermediate reasoning. Performance of LLMs in reasoning problems are further enhanced by providing in-context learning by means of few shot prompting. Here, we propose a pipeline approach to improve numerical reasoning capability of LLM on FinQA dataset. The first module of our framework consists a retrieval model to extract relevant facts for each question using LORA Flan-T5. Following this we propose program of thought prompting using these retrieved facts in a few shot setting where in-context examples are selected using a policy gradient method.

## REFERENCES

- [1] Subhendu Khatuya, Rajdeep Mukherjee, A Ghosh, M Hegde, K Dasgupta, N Ganguly, S Ghosh, and P Goyal. 2024. Parameter-Efficient Instruction Tuning of Large Language Models For Extreme Financial Numeral Labelling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Mexico City, Mexico.
- [2] Subhendu Khatuya, K Sinha, S Ghosh, N Ganguly, and P Goyal. 2024. Instruction-Guided Bullet Point Summarization of Long Financial Earnings Call Transcripts. In *Association for Computational Linguistics*. ACM, Washington, DC, USA.
- [3] 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*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 3550–3561. <https://doi.org/10.18653/v1/2023.findings-acl.219>

<sup>1</sup>Our dataset and codes are publicly available at <https://github.com/subhendukhatuya/FLAN-FinXC>

<sup>2</sup>Publicly available ECTSum Dataset: <https://github.com/rajdeep345/ECTSum>

<sup>3</sup>We make the codebase publicly available at <https://github.com/subhendukhatuya/FLAN-FinBPS>