

Auto-Generating Earnings Report Analysis via an Augmented LLM

Van-Duc Le

Seoul National University, Seoul, South Korea
levanduc@snu.ac.kr

Abstract. Financial analysis heavily relies on the evaluation of earnings reports to gain insights into company performance. Traditional generation of these reports requires extensive financial expertise and is time-consuming. With the impressive progress in Large Language Models (LLMs), a wide variety of financially focused LLMs has emerged, addressing tasks like sentiment analysis and entity recognition in the financial domain. This paper presents a novel challenge: developing an LLM specifically for automating the generation of earnings reports analysis. Our methodology involves an in-depth analysis of existing earnings reports followed by a unique approach to fine-tune an LLM for this purpose. This approach combines retrieval augmentation and the generation of instruction-based data, specifically tailored for the financial sector, to enhance the LLM’s performance. With extensive financial documents, we construct financial instruction data, enabling the refined adaptation of our LLM to financial contexts. Preliminary results indicate that our augmented LLM outperforms general open-source models and rivals commercial counterparts like GPT-3.5 in financial applications. Our research paves the way for streamlined and insightful automation in financial report generation, marking a significant stride in the field of financial analysis.

Keywords: LLMs · financial analysis · instruction tuning.

1 Introduction

Large Language Models (LLMs) have shown remarkable proficiency in diverse applications, including financial analysis. BloombergGPT [7], a specialized LLM for finance, is built from scratch and excels in various financial tasks, though it is costly due to its use of proprietary data and training expenses of over \$2 million. Conversely, FinGPT [8] is a low-cost, open-source LLM fine-tuned on instructional data, focusing on basic financial tasks like sentiment classification and named entity recognition. This paper explores more complex financial challenges, particularly earnings report analysis.

A company’s earnings report serves as a comprehensive self-declaration regarding its financial performance within a given fiscal period. Consequently,

earnings analysis assumes a pivotal role in financial assessment, aiming to garner a nuanced understanding of the company’s performance. Therefore, earnings report analysis necessitates adept handling of various financial intricacies, including considerations such as the peer sector dynamics, challenges related to calendarization, and the nuanced language employed in the executives’ communications. In the next section, we provide details about how earnings analysis can be conducted via our augmented system.

Instruction tuning, as proposed by [5], stands out as a cost-effective and efficient technique for aligning an LLM with a specific downstream task, such as the earnings analysis. The process involves the collection of instruction-following data, subsequently used to fine-tune the LLM. Since human generation datasets are very costly, a pragmatic alternative involves initiating the fine-tuning process with seed examples and employing a high-capacity teacher LLM such as OpenAI GPT-3.5 to generate a diverse set of instructional data. While this approach has demonstrated success across various problem domains, such as general question-answering [5] and visual understanding [4], its applicability to earnings report analysis remains an open question.

In addressing domain-specific challenges, an LLM frequently leverages the retrieval augmented generation (RAG) approach [3], a methodology that furnishes the model with domain-specific context. In our work, aimed at grounding generation instructions within the financial domain, we introduce a novel method termed **retrieval-augmented instruction data generation**. This methodology integrates both the financial domain context and the instructional prompt during data collection, guiding the teacher LLM in generating questions and answers aligned with the provided context. The approach yields a rich set of financial instruction-following data, important for the fine-tuning of a financial-augmented LLM.

In summary, this research makes several significant contributions. Firstly, we address the earnings report analysis problem, representing a pivotal extension in the realm of LLMs focused on finance. Our novel approach involves a retrieval-augmented instruction data generation method tailored for domain-specific tasks. This method generates instruction data contextually grounded in the financial domain. By fine-tuning an LLM on the curated instruction data, we demonstrate superior performance within the financial domain, much better than the well-known open-source LLM, Llama-2-7b [6], and comparable with the commercial GPT-3.5 model.

2 Earnings Report Analysis

Properly analyzing an earnings report of a company is a crucial task for any analyst because it involves a comprehensive understanding of the company’s financial health and performance. A few guidelines are provided on the web as to how to perform the analysis, but there is no definite standard on what should be included in the earnings report. Moreover, the analysis involves not just an objective evaluation of the numbers but also a subjective choice of which finan-

cial numbers to look for from a company’s earnings reports. Those components are what we call the Key Performance Indicators (KPIs). There are many components that make up the earnings analysis, but KPIs are one of the most basic and important numbers that the analyst chooses.

The KPIs are what we will use as the main content for auto-generating the analysis report. It is important to note that no KPIs are the same for each company and for each analyst. During the earnings call, the CEO usually chooses some set of numbers to argue that the company is doing great. But, the problem is that those numbers are chosen differently over time. Additionally, from the analyst’s point of view, he may not look for those numbers but rather needs the numbers from his own set of KPIs. These together raise discrepancies in the company’s analysis.

To address that discrepancy, we introduce a number of baseline KPIs that the analyst can use to initially gauge the company’s financial performance. Then, the analyst can manually update them to customize the analysis. The report will then be generated based on the results of those KPIs. In other words, it involves human-in-the-loop feedback from the analyst, which then prompts the model to auto-generate the analysis report. Adopting KPIs into our report generation process preserves the uniqueness of the analyst’s point of view while maintaining the consistency of the process. In the experiment part, we evaluate our model with basic KPI-based questions like the quarterly dividend.

3 Retrieval Augmented Instruction Tuning

This section describes our retrieval-augmented instruction tuning method and how to build a financially augmented LLM by fine-tuning from general financial instruction data.

3.1 General financial instruction data

Table 1. Instruction prompt to generate general financial instruction data and a data sample of generated instruction-following data. $\{num_questions_per_chunk\}$ is the number of data samples we want to generate for a chunk of text.

Instruction: You are an earnings report analyst. Your task is to ask $\{num_questions_per_chunk\}$ questions to understand a company, its financial report, and its key financial performance. The questions should be diverse in nature across the document. Restrict the questions to the context of the information provided.
Question: What is the fiscal year-end date for NVIDIA Corporation?
Answer: The fiscal year end date for NVIDIA Corporation is January 28.

To fine-tune a financially augmented LLM, we need general financial instruction data that can be generated from the context of financial documents, such

as quarterly reports. This process entails initially segmenting the financial document into discrete text chunks. Subsequently, each text chunk serves as the context for the generation of instruction-following data pertaining to financial nuances. Employing a teacher LLM, such as GPT-3.5-turbo, we extract insights and generate relevant instructions within the financial context. The specifics of this process, along with an illustrative example, are delineated in table 1. We opt for a configuration with *num_questions_per_chunk* set to 10, amassing a dataset of 400 financially augmented instruction instances derived from a single financial report.

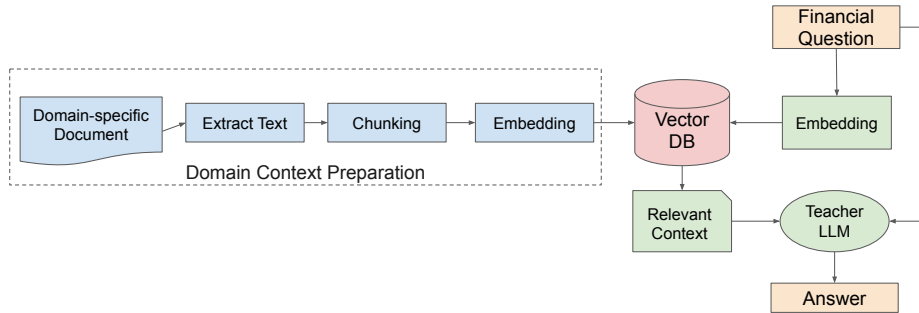


Fig. 1. The retrieval augmented instruction data generation workflow.

In Fig. 1, we present the workflow of our innovative retrieval-augmented instruction data generation methodology. Commencing with a domain-specific document indexing step, we populate a vector database (such as ChromaDB) to establish a contextual foundation within the domain. Subsequently, the instruction prompt undergoes processing by the teacher LLM (such as GPT-3.5), which generates pertinent question-answer pairs based on the extracted contextual information. This comprehensive approach ensures that the instruction generation is intricately tied to the relevant domain context, enhancing the quality and specificity of the resultant data.

3.2 Earnings analysis seed instructions

Precise guidance for the teacher LLM in the form of seed instructions is essential to ensure the generation of accurate and varied instruction-following data. Given the intricacies involved in earnings analysis, we employ conversation-based instruction prompts tailored to facilitate the analysis of a company over a specific period. This contextual awareness enables the model to comprehend the entire analysis methodology, contributing to its ability to generate coherent and insightful earnings analysis. A comprehensive breakdown of various instruction types, accompanied by illustrative prompt examples and their corresponding financial documents, is presented in table 2.

Table 2. Seed instructions for earnings analysis instruction-following data generation.

No.	Instruction Type	Prompt Example	Relevant Documents
1	Company core information	What is Nvidia and its business sector?	Press Release
2	Key financial indicators	What is the revenue?	Press Release, Earnings Report
3	Comparison	Can you compare the revenue of Nvidia with its peer group in the semiconductors industry?	Press Release, Earnings Report
4	Outlook	Can you give the outlook for the revenue of Nvidia in the next quarter?	Press Release, Earnings Report
5	Summary	Can you summarize the earnings report and executives' statements of Nvidia?	Press Release, Earnings Call Transcript
6	Analysis	Given the information above, can you generate an earnings report analysis for Nvidia in this quarter?	Equity Research Report

3.3 Augmented LLM

Our primary goal is to harness the capabilities of a pre-trained LLM in conjunction with financially augmented instruction data. In this work, we adopt Llama-2-7b [6] as the foundational pre-trained LLM and employ LoRA [2] method for efficient parameter-specific fine-tuning. To enhance memory efficiency and expedite the training process, we incorporate a 4-bit compression technique following the QLoRA method [1]. The training data has the following format, in which *sample* is a training sample of the financial instruction dataset.

We have provided context information below.

```
-----
{sample['context']}
```

Given this information, please answer the question:

```
{sample['query']}
```

```
Answer: {sample['answer']}
```

4 Experimental results

4.1 Experiment set up

We employed the Google Cloud virtual machine for training on four Tesla T4 GPUs, each equipped with 16GB of RAM. The training process spanned three epochs, maintaining a consistent learning rate of $2e-4$ while adhering to a maximum sequence length of 2048 for optimal model performance. The total training time was 40 hours.

We train our model on the financial instruction dataset generated on earnings reports from two companies in the semiconductor sector, Nvidia and AMD, in the third quarter of 2023 and evaluate the earnings report of Broadcom Inc. in the same period. In total, we have 800 training samples of financial instruction data.

4.2 Evaluation metrics

In this section, we present both quantitative and qualitative outcomes from our experimentation with a financial-augmented LLM using samples of earnings analysis question-answer (QA). To benchmark our model’s performance, we compared it against two baselines: Llama-2, a general open-source LLM from Meta [6], and GPT-3.5, a commercial LLM from OpenAI. All three models were equipped with the same financial contexts and questions. Our evaluation leverages two key metrics: correctness and semantic similarity. Correctness assesses the model’s ability to provide accurate answers to the given questions. As both the generated and ground-truth answers are in natural language, we define the evaluation scores on a scale from 1 to 10 and employ GPT-4 as the reference for judgment since GPT-4 can provide human-equivalent evaluation. Semantic similarity gauges how closely the generated responses align with the base answers, with lower values indicating better alignment.

4.3 Quantitative comparison

In this section, we offer a quantitative analysis of our model in comparison to two baseline models, employing the aforementioned evaluation metrics of correctness and semantic similarity. The quantitative results are displayed in table 3, unequivocally demonstrating the superior performance of our augmented model in the financial domain when contrasted with the open-source Llama-2. Remarkably, our model achieves comparable results to the commercial model, GPT-3.5. These findings underscore the efficacy and competitiveness of our approach, positioning it as a noteworthy advancement in the field.

Table 3. Quantitative comparison of our Financially-augmented LLM with Llama-2-7b and GPT-3.5.

Model	Correctness (1: worst, 10: best)	Semantic Similarity (smaller is better)
Financially-augmented LLM	4.6	0.14427
Llama-2-7b	2.8	0.19126
GPT-3.5	5.3	0.10659

Table 4. Qualitative comparison of the examples of earnings analysis for our augmented LLM, Llama-2-7b and GPT-3.5.

<p>Our Financially-augmented LLM</p> <p><i>Question:</i> What is Broadcom Inc.?</p> <p><i>Generated Answer:</i> Broadcom Inc. is a technology company that designs, develops, and supplies semiconductor and infrastructure software solutions.</p> <p><i>Question:</i> What is the company’s quarterly dividend for the third quarter of Fiscal Year 2023?</p> <p><i>Generated Answer:</i> 4.60</p> <p><i>Ground-truth Answer:</i> The company’s quarterly dividend for the third quarter of Fiscal Year 2023 is \$4.60 per share.</p>
<p>Llama-2-7b</p> <p><i>Question:</i> What is Broadcom Inc.?</p> <p><i>Generated Answer:</i> 1. An American manufacturer of semiconductor devices and infrastructure software solutions. 2. A software company that provides solutions for managing applications.</p> <p><i>Question:</i> What is the company’s quarterly dividend for the third quarter of Fiscal Year 2023?</p> <p><i>Generated Answer:</i> 368.40</p>
<p>GPT-3.5</p> <p><i>Question:</i> What is Broadcom Inc.?</p> <p><i>Generated Answer:</i> Broadcom Inc. is a global technology leader that designs, develops, and supplies semiconductor and infrastructure software solutions.</p> <p><i>Question:</i> What is the company’s quarterly dividend for the third quarter of Fiscal Year 2023?</p> <p><i>Generated Answer:</i> The company’s quarterly dividend for the third quarter of Fiscal Year 2023 is \$4.60.</p>

4.4 Qualitative comparison

We conducted a rigorous qualitative comparison of our augmented model with two baseline models. To ensure a fair evaluation, all three models were provided with the same quarterly report document, and we employed an identical embedding model for vector indexing. Table 4 showcases an illustrative example of a question pertaining to earnings analysis, along with the generated answers from each model. The qualitative analysis presented in the table distinctly demonstrates that our Financially-augmented model consistently produces answers

that are notably contextually relevant to the given question when compared to Llama-2, and similar to GPT-3.5 in this regard. This empirical evidence underscores the efficacy and superior performance of our model within the specialized domain of financial analysis.

5 Conclusion

This research paper aims to apply the field of large language models to the financial domain, with a particular focus on addressing the challenges of analyzing earnings reports. The introduction underscores the significance of this extension for researchers investigating the capabilities of large language models in the financial sector. The paper introduces an innovative approach centered around a retrieval-augmented instruction data generation method tailored specifically for tasks within the financial domain. The results highlight the efficacy of the augmented model, showcasing its potential to democratize the process of earnings report analysis.

References

1. Detrmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314 (2023)
2. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
3. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* **33**, 9459–9474 (2020)
4. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
5. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. <https://crfm.stanford.edu/2023/03/13/alpaca.html> **3**(6), 7 (2023)
6. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
7. Wu, S., Irsoy, O., Lu, S., Dabrovolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G.: Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564 (2023)
8. Yang, H., Liu, X.Y., Wang, C.D.: Fingpt: Open-source financial large language models. arXiv preprint arXiv:2306.06031 (2023)