

**Article info****Type of article:**

Original research paper

**DOI:**

<https://doi.org/10.58845/jstt.utt.2026.en.6.2.218-231>

**\*Corresponding author:**

Email address:

[thienth@utt.edu.vn](mailto:thienth@utt.edu.vn)

**Received:** 11/09/2025**Received in Revised Form:**

09/11/2025

**Accepted:** 16/12/2025

## RAG-IT: Retrieval-Augmented Instruction Tuning for Automated Financial Analysis- A Case Study for the Semiconductor Sector

Van-Duc Le<sup>1</sup>, Tien-Cuong Bui<sup>2</sup>, Hai-Thien To<sup>3</sup><sup>1</sup>Independent Researcher, Seoul, South Korea<sup>2</sup>Arontier Co., Ltd, Seoul, South Korea<sup>3</sup>University of Transport Technology, Hanoi, Vietnam

**Abstract:** Financial analysis relies heavily on the interpretation of earnings reports to assess company performance and guide decision-making. Traditional methods for generating such analyses require significant financial expertise and are often time-consuming. With the rapid advancement of Large Language Models (LLMs), domain-specific adaptations have emerged for financial tasks such as sentiment analysis and entity recognition. This paper introduces RAG-IT (Retrieval-Augmented Instruction Tuning), a novel framework designed to automate the generation of earnings report analysis through an LLM fine-tuned specifically for the financial domain. Our approach integrates retrieval augmentation with instruction-based fine-tuning to enhance factual accuracy, contextual relevance, and domain adaptability. We construct a sector-specific financial instruction dataset derived from semiconductor industry documents to guide the LLM adaptation to specialized financial reasoning. Using NVIDIA, AMD, and Broadcom as representative companies, our case study demonstrates that RAG-IT substantially improves a general-purpose open-source LLM and achieves performance comparable to commercial systems like GPT-3.5 on financial report generation tasks. This research highlights the potential of retrieval-augmented instruction tuning to streamline and elevate financial analysis automation, advancing the broader field of intelligent financial reporting.

**Keywords:** LLMs, financial analysis, instruction tuning.

### 1. Introduction

Large Language Models (LLMs) have shown remarkable proficiency in diverse applications, including financial analysis. BloombergGPT [1], 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 [2] is an open-source low-cost 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 [3], 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 [3] 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 [5], a methodology that provides 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 designed for domain-specific tasks. This method generates instruction data contextually grounded in the financial domain. By fine-tuning an LLM with 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. Structured Components of Earnings Report Analysis

A comprehensive earnings report analysis goes beyond the computation of financial metrics; it involves a structured reasoning process across several inter-connected dimensions of company performance. To ensure both interpretability and reproducibility of the analysis, we decompose the process into a series of analysis modules, each corresponding to a specific section of an earning report. These modules form the basis of our data-generation and report-synthesis framework.

Each instruction type encapsulates a distinct analytical objective, as detailed below. To ensure that our six-component structure is not arbitrary, we consulted two financial experts with MBA degrees and many years of experience in large corporations. Their feedback confirmed that this decomposition aligns well with the workflow used in professional earnings reports, providing practical grounding for our design choices.

### (1) Company Core Information

This section introduces the company's background, business model, and sectoral context. It provides foundational context for the subsequent financial and strategic assessments. Typical questions at this level include identifying the company's primary business domains, product lines, and target markets. The main information sources are the company's press releases and official websites, which ensure factual accuracy. This step grounds the report-generation process,

enabling the model to contextualize later KPI-based interpretations within the company's operational scope.

#### (2) Key Financial Indicators (KPIs)

The next stage involves extracting and interpreting core quantitative metrics that define the company's financial performance. Examples include revenue, net income, operating margin, earnings per share (EPS), and cashflow. These KPIs are retrieved from earnings reports and press releases, then aligned with the analyst's custom-defined KPI set. This allows the analysis to be both standardized and flexible-standardized in structure, yet personalized to reflect the analyst's expertise. In our framework, these KPIs serve as anchor points for the model's narrative generation, allowing the auto-generated analysis to maintain factual grounding while adapting to subjective analytical focus. Indeed, financial KPIs have been characterized as metrics that provide insights into underlying business performance across categories such as profitability, liquidity, efficiency, valuation and leverage. [7, 8]

#### (3) Comparative Analysis

To provide market context, comparative analysis benchmarks the company's performance against peers in the same industry. This may involve ratio-based comparisons (e.g., revenue growth rate, profit margin) or relative ranking within the sector. The comparison step enhances interpretability by situating a company's performance within broader market trends. The relevant documents include both earnings reports of competitors and industry summaries, allowing the model to contextualize differences in business cycles or strategic focus.

#### (4) Outlook Analysis

Forward-looking analysis assesses management's guidance, analyst consensus, and projected financial trends. By drawing from press releases, earnings reports, and occasionally analyst forecasts, the model can synthesize expectations for future performance, such as

projected revenue growth or cost-optimization strategies. This component also reflects qualitative signals, such as management tone or sentiment extracted from the earnings call transcript, which complement the quantitative KPIs. Integrating these elements allows the model to generate a balanced view that combines data-driven projections with managerial perspectives.

#### (5) Summary and Executive Sentiment

The summary component synthesizes key insights from both financial and textual sources, such as executives' remarks and press announcements. This step aims to distill the overall narrative conveyed during the earnings call, highlighting management's confidence levels, strategic priorities, or risk acknowledgments. We use this component to train the model to capture soft indicators that are not directly observable in numerical data but are critical for evaluating corporate outlook and investor sentiment.

#### (6) Integrated Analysis Report

Finally, all the preceding elements are integrated into a cohesive earnings report analysis. This module combines objective KPI evaluation, comparative benchmarks, and qualitative outlook synthesis to generate a holistic view of the company's performance in the given quarter. The structure and tone are inspired by professional equity research reports, ensuring the generated output is aligned with financial reporting standards. The analyst remains in the loop by reviewing or updating the KPI set, thereby guiding the model's focus in subsequent iterations.

### 3. Financial Instruction Data

Developing a financially augmented large language model (LLM) requires a well-structured and context-rich dataset that reflects the language, reasoning patterns, and analytical conventions of financial reporting. To achieve this, we construct a two-tier dataset which serves as the foundation for instruction tuning and retrieval-augmented fine-tuning.

#### 3.1. General Financial Instruction Data

To fine-tune a financially augmented LLM, we first require a diverse set of general financial instruction data that reflect real-world financial document structures and semantics. Such data are generated from the contextual content of company financial disclosures, including quarterly and annual reports, management commentary, and earnings call summaries.

The data generation process begins by segmenting each financial document into discrete, semantically coherent text chunks. Each chunk represents a self-contained financial context that can independently support the generation of multiple instruction-following pairs. Examples of such contexts include a company’s identity, its industry classification, and reported key financial metrics. For instance, our segmentation of NVIDIA’s quarterly earnings report yielded chunks addressing queries such as:

- What is NVIDIA?
- What are the main industry sectors of NVIDIA?
- What is the revenue of NVIDIA for the second quarter of fiscal year 2024?
- What is the GAAP and Non-GAAP earnings per diluted share for the quarter?
- How many shares did NVIDIA repurchase during the second quarter of fiscal 2024, and for what amount?

Each chunk thus acts as a contextual unit from which the model can learn to map document-

grounded financial text to relevant natural language queries. Using a teacher LLM such as GPT-3.5-turbo, we automatically generate instruction–response pairs based on the textual content of each chunk. The teacher model is prompted to extract domain-specific insights, numerical facts, and relational patterns (e.g., quarter-over-quarter growth, year-over-year comparison, dividend schedules, or share repurchases).

To ensure balanced coverage, we set the configuration parameter `num_questions_per_chunk` to 10. For a single quarterly report containing approximately 40 well-defined chunks, this setup yields around 400 financial instruction–response pairs. These pairs collectively form the general financial instruction dataset, which serves as the base layer for fine-tuning the financially augmented LLM.

The resulting dataset encompasses multiple dimensions of financial understanding, including:

1. Entity-level understanding: Identifying the company, fiscal period, and report structure.
2. Numerical reasoning: Computing and interpreting KPIs such as revenue, operating income, and EPS.
3. Temporal reasoning: Comparing performance across quarters and fiscal years.
4. Policy and governance insight: Understanding board authorizations, dividend policies, and share repurchase programs.

**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

---

<b>Instruction:</b>	You are an earnings report analyst. Your task is to ask <code>num_questions_per_chunk</code> 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.
<b>Question:</b>	What is the fiscal year-end date for NVIDIA Corporation?
<b>Answer:</b>	The fiscal year end date for NVIDIA Corporation is January 28.

---

This structured data generation process enables the LLM to capture the linguistic and

analytical characteristics of financial narratives, providing a strong foundation for subsequent

retrieval-augmented fine-tuning. An illustrative sample of the generated instruction-following data is shown in Table 1.

In summary, the general financial instruction dataset serves two purposes: (1) it provides a systematic foundation of financial question–answer pairs derived directly from authentic corporate disclosures, and (2) it establishes the contextual grounding necessary for extending the LLM’s capacity to handle retrieval-augmented financial reasoning tasks discussed in the next section.

**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, diverse, and contextually grounded instruction-following data. Given the complexities inherent in financial text, particularly in earnings reports, the model must understand not only numerical facts but also qualitative relationships, comparative metrics, and

managerial sentiment. To this end, we design a structured set of seed instructions that guide the teacher model to produce high-quality financial reasoning data.

The seed instructions are crafted to emulate the reasoning process of human financial analysts. They cover multiple analytical dimensions, including factual extraction, comparison, temporal reasoning, and interpretive commentary. We employ conversation-based instruction prompts to create a natural flow of inquiry-mimicking how an analyst might sequentially explore a company’s quarterly performance. This contextual chaining enables the teacher model to capture dependencies between questions and answers, resulting in coherent and analytically rich datasets.

Table 2 presents a taxonomy of six instruction types with representative examples and their associated document sources. Each instruction type plays a distinct role in the generation of financial instruction-following data:

**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

• **Company Core Information:** Focuses on the company’s identity, business model, and operational domain. These questions establish the foundational context of the analysis.

• **Key Financial Indicators:** Targets quantitative performance metrics such as revenue, earnings per share (EPS), or operating margin. The responses typically include both the reported

values and brief interpretations.

- **Comparison:** Involves benchmarking company metrics against peer firms or historical performance to capture relative strengths or weaknesses.

- **Outlook:** Encapsulates management guidance, analyst consensus, and forward-looking statements that indicate projected financial performance. These prompts encourage narrative synthesis beyond raw numbers.

- **Summary:** Integrates insights from multiple sections of the earnings release and call transcripts, summarizing the quarter's highlights and strategic focus.

- **Analysis:** Represents the highest level of reasoning, combining numerical interpretation, strategic evaluation, and investor sentiment to form a holistic analysis. This aligns most closely with the structure of an equity research report.

Each instruction type is linked to specific source documents—such as press releases, earnings reports, call transcripts, or equity research summaries—to ensure grounding in verified financial text. During instruction data generation, the teacher LLM is conditioned on both the prompt and its associated document snippet, allowing the resulting responses to be evidence-based and contextually aligned.

Furthermore, we design the prompting sequence to include interdependent instructions, where later prompts (e.g., Outlook or Analysis) depend on information extracted in earlier steps (e.g., Key Financial Indicators). This multi-turn structure improves the model's capacity to integrate cross-referenced information, mirroring the analytical reasoning process of human experts.

The seed instructions thus serve as a blueprint for generating a comprehensive and multi-dimensional financial dataset. By training the student model on these structured instruction–response pairs, the LLM learns to:

1. Accurately extract and interpret quantitative metrics from textual disclosures.

2. Contextualize these metrics within industry and temporal frames.

3. Generate coherent narratives that explain the causes and implications of financial outcomes.

In summary, earnings analysis seed instructions establish a consistent and interpretable framework for generating high-quality financial reasoning data. They ensure that the model learns not only to retrieve factual content but also to perform structured analysis—an essential step toward developing a retrieval-augmented financial LLM capable of producing explainable and analyst-like insights.

#### **4. Retrieval-Augmented Instruction Tuning Method**

This section describes our retrieval-augmented instruction tuning method and outlines how we construct a financially augmented large language model (LLM) through fine-tuning using general financial instruction data and retrieval-augmented financial reasoning data. The overall framework aims to enhance an LLM's ability to reason about financial information in context, improving its performance on document-grounded financial tasks.

In Fig. 1, we present the workflow of our retrieval-augmented instruction data generation and model fine-tuning methodology. The approach is designed to tightly couple context retrieval, instruction generation, and parameter-efficient fine-tuning. The method consists of two main components: (1) the retrieval-augmented instruction data generation process, and (2) the financially augmented LLM fine-tuning process.

##### **4.1. Retrieval-Augmented Instruction Data Generation**

This component focuses on creating context-grounded, instruction-following financial data derived from authentic financial documents. The goal is to generate diverse and accurate question–answer pairs that reflect the reasoning tasks encountered in financial analysis.

The pipeline begins with a document

indexing and retrieval stage. Financial documents such as quarterly earnings reports, press releases, and SEC filings are segmented into coherent text chunks. These chunks are then embedded using a dense embedding model and stored in a vector database (e.g., ChromaDB), which supports semantic retrieval.

Given a financial topic or company name, the system retrieves relevant chunks based on similarity search. These chunks are passed as contextual input to a teacher model (e.g., GPT-3.5-turbo), which generates instruction–response pairs grounded in the retrieved financial context. Each generated sample follows the structure:

```
{
  "instruction": <financial query>,
  "input": <retrieved document context>,
  "output": <teacher-generated answer>
}
```

This process ensures that generated data remain faithful to real financial statements and reduce hallucinations by grounding responses in retrieved evidence. In addition, conversational-style prompting is employed to elicit richer, more analytical responses that capture patterns such as revenue growth, expense variation, or year-over-year comparisons.

After generation, the data undergo normalization (e.g., numeric unit alignment, temporal standardization, duplicate removal). The final dataset integrates both general and analytic instructions, supporting a wide range of reasoning patterns that align with financial interpretation and decision-making.

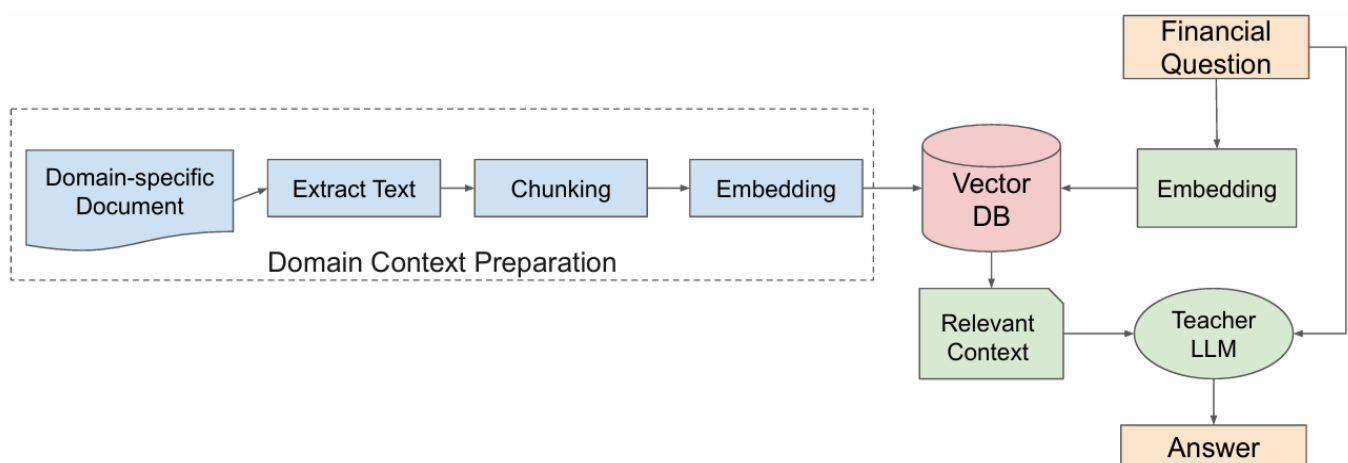


Fig. 1. The retrieval-augmented instruction data generation workflow

#### 4.2. Financially Augmented LLM Fine-Tuning

Once the retrieval-augmented financial instruction dataset is prepared, we fine-tune a pre-trained LLM to develop a financially specialized reasoning capability. The fine-tuning procedure focuses on efficiency and stability while preserving general language understanding.

During training, each instruction sample (sample) follows the previously described schema, where the instruction represents a financial query, the input provides relevant context, and the output serves as the correct answer. The model learns to interpret financial context, retrieve relevant details,

and synthesize grounded analytical conclusions.

This approach ensures that the fine-tuned LLM can:

- Perform context-grounded financial question answering with factual precision.
- Conduct multi-hop reasoning, such as temporal comparisons or cross-metric correlations.
- Generate human-like financial commentary aligned with analyst-style interpretations.

In summary, the retrieval-augmented instruction tuning framework integrates contextual data generation and parameter-efficient fine-tuning to build a domain-adapted financial LLM. The

retrieval process ensures factual grounding, while instruction tuning refines the model's analytical expressiveness. Together, these two components yield a system capable of delivering accurate, interpretable, and contextually grounded financial reasoning.

## 5. Experimental results

### 5.1. Experiment configurations and set up

We employ Llama-2-7B [6] as the base model and fine-tune it using the Low-Rank Adaptation (LoRA) method [9]. Following the QLoRA setup [10], we enable 4-bit (NF4) quantization using `prepare_model_for_kbit_training`. Our LoRA configuration uses a rank of 64, `lora_alpha = 16`, and a dropout of 0.1, applied to all attention and feed-forward projection layers with no bias adaptation. Fine-tuning is conducted using SFTTrainer with a maximum sequence length of 1024 tokens and packing enabled to improve training throughput.

We train for 3 epochs with a global batch size of 8 (per-device batch size 2, gradient accumulation 2) using the `paged_adamw_32bit` optimizer. The learning rate is set to  $2e-5$  with a constant schedule and a warmup ratio of 0.03. We disable gradient checkpointing and use `ddp_find_unused_parameters=False` for efficient multi-GPU execution. The maximum gradient norm is clipped to 0.3.

We employed the Google Cloud virtual machine for training on four Tesla T4 GPUs, each equipped with 16GB of RAM. Due to the limited compute capability and memory bandwidth of T4 GPUs combined with 4-bit QLoRA dequantization overhead and activation packing, the end-to-end fine-tuning of our 800-sample dataset requires approximately 40 hours.

For the retrieval component, we use the text-embedding-3-large encoder to embed documents. The corpus is chunked into 512-token segments with a 128-token overlap. At inference time, the retriever selects the top  $k=3$  most relevant chunks

using cosine similarity.

To ensure the correctness of GPT-3.5-generated training data, we apply a two-stage quality control process: (1) rule-based filtering (format validation, length constraints, and inconsistency checks), and (2) manual inspection of a randomly sampled 15% subset to identify hallucinations, numerical inconsistencies, and unclear instructions. Problematic items are revised or removed before final training.

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.

### 5.2. 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.

The prompting for GPT-4 judgment is shown in the listing 1.

Listing 1: GPT-4 Judge Prompt

You are an expert evaluator. Your task is to

compare a candidate answer against a reference answer and assign a numeric score based on the scoring rubric provided. You must follow the rubric strictly and output only the final score as an integer from 1 to 10.

Score 1: The answer is completely unrelated to the reference.

Score 3: The answer has minor relevance but does not align with the reference.

Score 5: The answer has moderate relevance but contains inaccuracies.

Score 7: The answer aligns with the reference but has minor errors.

Score 9: The numbers match perfectly, structure is similar.

Score 10: Completely accurate and perfectly aligned.

**5.3. Quantitative comparison**

**Table 3.** Quantitative comparison of our Financially-augmented LLM with Llama-2-7b and GPT-3.5. Correctness is in [1,10], where 1 is the worst and 10 is the best. For similarity score, smaller is better

Model	Correctness	Semantic Similarity
Financially-augmented LLM	4.6	0.14427
Llama-2-7b	2.8	0.19126
GPT-3.5	5.3	0.10659

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.

**5.4. 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 presents examples of questions related to earnings analysis, along with the corresponding generated answers from each model. The qualitative analysis reveals clear distinctions in the quality and contextual relevance of responses.

In the general company information question, all three models, the open-source general-purpose

Llama-2-7b, our Financially-augmented LLM, and the commercial GPT-3.5, produce correct and sufficiently detailed descriptions of Broadcom Inc. This indicates that broad corporate identity queries can be reliably answered even without financial-domain specialization. However, in the financial metric extraction query, our model provides a more accurate and semantically aligned answer than the open-source baseline and shows a level of correctness that closely matches the response from the powerful commercial GPT-3.5 model.

Our Financially-Augmented LLM. Our model consistently demonstrates high contextual accuracy and precise numerical reasoning. It is important to note that our evaluation framework prioritizes numeric correctness over linguistic completeness. A prime example is the dividend question, where the generated response (“4.60”) directly matches the ground-truth value without superfluous phrasing, preserving factual consistency with the source document. This demonstrates that the model effectively retrieves and integrates relevant context from financial reports. The retrieval-augmented instruction tuning process evidently allows the model to internalize document structure and financial semantics, producing domain-appropriate terminology (e.g.,

“semiconductor and infrastructure software solutions”) comparable to human financial analysts.

Llama-2-7b Baseline. The baseline Llama-2-7b model exhibits limited grounding and weaker numerical reasoning. Although it provides a generally correct company description, the responses show redundancy and lack of focus, as

evidenced by the listing of multiple partial statements. For numerical queries, the model frequently produces spurious values (e.g., "368.40"), demonstrating its inability to extract and reason over financial figures from context. These errors highlight the lack of document-grounded alignment and domain-specific adaptation in the base LLM.

**Table 4.** Qualitative comparison of model responses across different financial question types

(a) Question Type: General Company Information

Question: What is Broadcom Inc.?

Model	Generated Answer
Our Financially-augmented LLM	Broadcom Inc. is a technology company that designs, develops, and supplies semiconductor and infrastructure software solutions.
Llama-2-7b	1. An American manufacturer of semiconductor devices and infrastructure software solutions. 2. A software company that provides solutions for managing applications.
GPT-3.5	Broadcom Inc. is a global technology leader that designs, develops, and supplies semiconductor and infrastructure software solutions.

(b) Question Type: Financial Metric Extraction Query

Question: What is the company’s quarterly dividend for the third quarter of Fiscal Year 2023?

Ground-truth: \$4.60 per share.

Model	Generated Answer
Our Financially-augmented LLM	\$4.60
Llama-2-7b	\$368.40
GPT-3.5	The company’s quarterly dividend for the third quarter of Fiscal Year 2023 is \$4.60.

GPT-3.5 Teacher Model. GPT-3.5 generates highly coherent and semantically accurate responses, often matching or exceeding the clarity of human-written financial summaries. Its answer for the dividend question exactly reproduces the ground truth, confirming strong factual precision and contextual comprehension. However, GPT-3.5 relies heavily on its large-scale pretrained knowledge and less on explicit document retrieval, occasionally leading to generic phrasing or assumptions when context is ambiguous. Despite this, it serves as an effective teacher model for generating context-grounded instruction data due to its consistent reasoning quality.

Overall Comparison. Overall, our Financially-Augmented LLM achieves a balance between GPT-3.5’s reasoning depth and Llama-2’s fine-tuning adaptability. The model effectively integrates retrieval context to ensure factual correctness while maintaining natural linguistic fluency. Qualitatively, it shows a clear improvement in interpretability, numeric reliability, and contextual grounding compared to the baseline. This result substantiates the effectiveness of retrieval-augmented instruction tuning in enabling domain-specialized financial reasoning within smaller, efficient LLMs.

**5.5. More Success and Failure Cases**

**Table 5.** Representative success and failure cases of our model across various financial metric types

Question	Metric Type	Generated Answer	Judge Model Reasoning	Score
What is the company's quarterly dividend for the third quarter of Fiscal Year 2023?	dividend	The company did not declare or pay any dividends for the third quarter of Fiscal Year 2023.	The model states that no dividend was declared, contradicting the ground truth that a dividend was paid.	1
What is the reported revenue for the third quarter of Fiscal Year 2023?	revenue-related	The reported revenue for the third quarter of Fiscal Year 2023 is \$13,507 million, up 101% from the same period last year.	The model provides revenue of \$13,507M, while the correct figure is \$8,876M, indicating a significant numeric error.	5
Can you provide a breakdown of the revenue by business segment?	segment-level breakdowns	Based on the provided context, AMD's revenue by business segment for the three months ended July 1, 2023 and June 25, 2022 are as follows: * Data Center: \$1,321 million in net revenue for Q2 2023 and \$1,486 million in Q2 2022. * Client: \$998 million in net revenue for Q2 2023 and \$2,152 million in Q2 2022. * Gaming: \$1,581 million in net revenue for Q2 2023 and \$1,655 million in Q2 2022. * Embedded: \$1,459 million in net revenue for Q2 2023 and \$1,257 million in Q2 2022. Note that these figures are based on management's judgment and allocation of expenses to the reportable segments, and may not necessarily reflect actual profitability by segment.	The model fabricates a detailed AMD revenue breakdown even though the reference document does not provide such information.	1
What is the quarterly period ended date?	period identification	Based on the context provided, the quarterly period ended date is July 1, 2023.	The model answers July 1, 2023, but the correct period end date is July 30, 2023.	7
What is the revenue of NVIDIA for second quarter fiscal year 2024?	revenue-related	NVIDIA's revenue for the second quarter of fiscal year 2024 was \$13.51 billion, up 101% from the same period last year and up 88% sequentially.	The model correctly provides \$13.51B and correctly add the percentage increase.	10

We provide additional success and failure cases of our model across different question types in Table 5, including different key financial indicators including dividend queries, revenue-related metrics, period identification, and segment-level breakdowns. These failure cases span multiple companies (Broadcom, NVIDIA, and AMD), illustrating that errors may occur not only in numeric extraction but also in semantic alignment with the ground-truth context. For each example, we list the question, the financial metric type, the model's generated answer, and the judgment from the GPT-4 judge model.

## 6. Related Work

Large Language Models (LLMs) have been increasingly applied to financial analysis, with early domain-specific models like BloombergGPT [1] and FinGPT [2] demonstrating the value of financial adaptation. Recent studies extend this trend toward more sophisticated analytical tasks. Jadhav and Mirza [11] provide a comprehensive review of 84 studies on LLMs in equity markets, emphasizing their applications in forecasting and portfolio management. Regulatory perspectives, such as the ESMA–Alan Turing Institute report [12], further highlight the importance of responsible and verifiable adoption in finance.

Instruction tuning has emerged as an efficient strategy for aligning LLMs with domain-specific objectives. Han et al. [13] present a comprehensive survey of instruction tuning methods, emphasizing alignment-centric paradigms. This paper also shows that data quality and domain diversity critically influence downstream task performance. In finance, Hirano and Imajo [14] propose building instruction-tuned LLMs without explicit instruction datasets via continual pretraining and model merging, and Tanabe et al. [15] introduce JaFIn, a Japanese financial instruction dataset that enhances domain adaptability.

Retrieval-Augmented Generation (RAG)

further improves factual grounding and contextual accuracy in financial reasoning. Zeng [16] presents QuantMCP, a retrieval-grounded framework linking LLMs to verified financial data sources, while Wang et al. [17] evaluate RAG and fine-tuning approaches for financial-statement analysis, highlighting their complementary strengths. Despite these advances, the integration of retrieval augmentation and instruction tuning for automated earnings report analysis remains unexplored. This work introduces RAG-IT, a framework unifying retrieval-augmented instruction data generation and fine-tuning to address this gap in financial LLM research.

## 7. 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.

Nevertheless, our approach has several limitations. RAG-IT may struggle with highly nuanced questions that require deep domain expertise, multi-hop reasoning over various financial metrics, or precise numerical computation. Its accuracy also depends heavily on the completeness and quality of retrieved evidences; missing or poorly retrieved context can degrade performance. Inference requires running both retrieval and generation, which introduces additional latency and computational cost compared to standard LLM inference. These limitations highlight important directions for future

research, including more robust retrieval methods, improved reasoning capabilities, and more efficient architectures for financial analysis.

## References

- [1] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D.S. Rosenberg, G. Mann. (2023). BloombergGPT: A large language model for finance. *ArXiv*. <https://arxiv.org/abs/2303.17564>
- [2] H. Yang, X.-Y. Liu, C.D. Wang. (2023). FinGPT: Open-source financial large language models. *ArXiv*. <https://doi.org/10.48550/arXiv.2306.06031>
- [3] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T.B. Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>
- [4] H. Liu, C. Li, Q. Wu, Y.J. Lee. (2023). Visual instruction tuning. *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*. *ArXiv*. DOI:10.48550/arXiv.2304.08485
- [5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 793, 9459–9474. *ArXiv*. <https://doi.org/10.48550/arXiv.2005.11401>
- [6] H. Touvron, L. Martin, K.R. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *ArXiv*. <https://doi.org/10.48550/arXiv.2307.09288>
- [7] A. Twin. (2024). KPIs: What are key performance indicators? Types and examples. *Investopedia*. (Retrieved February 8, 2024).
- [8] R. Bichachi. (2025). 30 financial metrics and KPIs to measure success in 2025. *NetSuite Resource Articles*. (accessed: Nov 9, 2025).
- [9] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen. (2021). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*. *ArXiv*. <https://doi.org/10.48550/arXiv.2106.09685>
- [10] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer. (2023). QLORA: Efficient finetuning of quantized LLMs. *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, 441, 10088–10115.
- [11] A. Jadhav, V. Mirza. (2025). Large language models in equity markets: Applications, techniques, and insights. *Frontiers in Artificial Intelligence*, 8, 1608365. doi: 10.3389/frai.2025.1608365
- [12] The European Securities and Markets Authority, The Alan Turing Institute and the Institut Louis Bachelier. (2025). Leveraging Large Language Models in Finance: Pathways to Responsible Adoption. *Technical Report*.
- [13] X. Han, J. Yang, T. Wang, Z. Bi, X. Song, J. Hao, J. Song. (2025). Towards alignment-centric paradigm: A survey of instruction tuning in Large Language Models. *ArXiv*, [abs/2508.17184](https://arxiv.org/abs/2508.17184). DOI:10.48550/arXiv.2508.17184
- [14] M. Hirano, K. Imajo. (2024). The construction of instruction-tuned LLMs for finance without instruction data using continual pretraining and model merging. *ArXiv*. [arXiv:2409.19854](https://arxiv.org/abs/2409.19854). <https://doi.org/10.48550/arXiv.2409.19854>
- [15] K. Tanabe, M. Suzuki, H. Sakaji, I. Noda. (2024). JaFIn: Japanese financial instruction dataset. *ArXiv*. <https://arxiv.org/abs/2404.09260>
- [16] Y. Zeng. (2025). QuantMCP: Grounding Large Language Models in verifiable financial reality. *ArXiv*. <https://arxiv.org/abs/2506.06622>

- [17] X. Wang, M. Brorsson. (2025). Can large language model analyze financial statements well? *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal), 2025*, pp. 196–206.