Heaven’s Light is Our Guide

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

Exploring the Effectiveness of Large Language Models in Financial Question Answering: A Comparative Analysis

Author

Mondol Mridul Provakar

Roll No. 1803062

Department of Computer Science & Engineering

Rajshahi University of Engineering & Technology

Supervised by

Emrana Kabir Hashi

Assistant Professor

Department of Computer Science & Engineering

Rajshahi University of Engineering & Technology

ACKNOWLEDGEMENT

First of all, I would like to thank Almighty, for his grace and blessings and for providing me with the diligence and enthusiasm along the way to accomplishing my thesis work.

I also want to express my sincere gratitude, admiration, and heartfelt appreciation to my supervisor **Emrana Kabir Hashi**, Assistant Professor, Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi. Throughout the year, she has not only provided me with the technical instructions and documentation to complete the work, but she has also continuously encouraged me, offered me advice, assisted me, and cooperated sympathetically whenever she deemed necessary. Her constant support was the most successful tool that helped me to achieve my result. Whenever I was stuck in any complex problem or situation, she was there for me at any time of the day. Without her sincere care, this work has not materialized in the final form that it is now at present.

I am also grateful to respected **Prof. Dr. Md. Ali Hossain**, Head of the Department of Computer Science & Engineering and all the respective teachers of Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi for their valuable suggestions and inspirations from time to time.

Finally, I would like to convey my thanks to my parents, friends, and well-wishers for their true motivations and many helpful aids throughout this work.

October 31, 2024 Mondol Mridul Provakar

RUET, Rajshahi

Heaven’s Light is Our Guide

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

CERTIFICATE

This is to certify that this thesis report entitled ***“Exploring the Effectiveness of Large Language Models in Financial Question Answering: A Comparative Analysis ”*** *submitted by* ***Mondol Mridul Provakar, Roll:1803062*** *in partial fulfillment of the requirement for the award of the degree of Bachelor of Science in Department of Computer Science & Engineering of Rajshahi University of Engineering & Technology, Bangladesh is a record of the candidate own work carried out by him under my supervision. This thesis has not been submitted for the award of any other degree.*

Supervisor

Emrana Kabir Hashi
Assistant Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology

External Examiner

Dr. Md. Nazrul Islam Mondal
Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology

ABSTRACT

The financial sector is witnessing a profound transformation driven by the integration of Artificial Intelligence (AI) technologies, with Large Language Models (LLMs) emerging as a promising frontier for innovation. These advanced AI models, renowned for their proficiency in natural language processing (NLP) tasks, hold the potential to revolutionize financial analysis by facilitating efficient and accurate information retrieval. This research investigates the efficacy of LLMs in addressing financial question-answering tasks, focusing specifically on two state-of-the-art models: Llama2-7b by Meta and Gemma-7b by Google. Despite their established prowess in general NLP tasks, their suitability for domain-specific applications, such as financial question answering, necessitates further exploration. Employing a comprehensive evaluation approach encompassing zero-shot prompt engineering, few-shot prompt engineering, and supervised fine-tuning methodologies, this study assesses the performance of Llama2 and Gemma using key metrics including ROUGE-L, cosine similarity, and human evaluation. The preliminary findings reveal significant distinctions between the two models. Llama2 demonstrates a higher frequency of generated answers; however, it is prone to hallucinations, often producing incorrect or incomplete information. In contrast, Gemma's performance is notably inferior, struggling to respond accurately to most queries. These observations emphasize the ongoing need for research and development endeavours to improve LLMs' capabilities in answering financial questions. By offering valuable insights into the strengths and limitations of LLMs in addressing financial inquiries, this study adds to the broader exploration of LLMs in financial contexts. Furthermore, it emphasizes the importance of refining model performance and developing domain-specific training datasets to unlock the full potential of LLMs in the financial domain.

CONTENTS

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Introduction

The financial industry is rapidly adopting Artificial Intelligence (AI) technology to address increasing needs for accuracy, effectiveness, and creative solutions. This trend underscores a deliberate transition towards leveraging advanced computational methods to navigate intricate financial landscapes, enhance decision-making processes, and foster sustained growth (Goodell et al., 2021) [1]. Large language Models (LLMs), exemplified by OpenAI's GPT series, have recently made significant strides in natural language processing (NLP), marking a notable milestone in AI development. Empowered by enhanced computational capabilities and refined algorithms, LLMs excel in comprehending complex contexts, addressing inquiries, and generating written content. Their transformative potential, particularly within the banking industry, is undeniable (Chang et al.,2024) [2]. Predictive modeling, intricate decision-making processes, and extensive data analysis characterize finance. LLMs promise to be a valuable financial tool because they rapidly process vast textual data. For instance, LLMs aid investors in making well-informed investment choices, assessing risks, and gaining insights into market trends through the analysis of financial data, market updates, and investor communications. Additionally, LLMs can understand natural language queries and provide prompt financial guidance, marking a significant advancement for the financial services sector (Zhao et al., 2024)[3].

1.2 Large Language Models in Financial Document Analysis

Large Language Models (LLMs) have showcased state-of-the-art performance across a spectrum of Natural Language Processing (NLP) benchmarks, encompassing evaluations like the Stanford Question Answering Dataset (SQuAD), TriviaQA, and Natural Questions. The evolution of LLMs has brought about pioneering methodologies including pre-training and fine-tuning, masked language modeling, permutation language modeling, few-shot learning, and more recently, zero-shot learning (Merkus et al., 2023)[4]. Complexities abound within the intricate landscape of financial documents, ranging from intricate regulatory filings to detailed financial statements and market reports. These documents often contain nuanced language, technical jargon, and contextual intricacies that pose significant challenges for traditional computational approaches. Large Language Models (LLMs) emerge as a transformative solution, providing unmatched capabilities in processing extensive amounts of textual data with exceptional speed and accuracy. By leveraging advanced techniques such as pre-training and fine-tuning, LLMs can effectively navigate complex financial documents, extract key insights, identify trends, and generate actionable recommendations. Moreover, their ability to understand and interpret natural language queries enables seamless interaction with financial data, empowering stakeholders to make informed decisions promptly. In essence, LLMs serve as a cornerstone in financial document analysis, offering a powerful toolset for deciphering the intricacies of complex financial documents and unlocking valuable insights hidden within. Their integration into financial workflows promises to streamline processes, enhance decision-making, and drive innovation within the financial industry.

1.3 Evaluation of LLMs in the Financial Domain

Assessing the performance of different Large Language Models (LLMs) within the financial sector is crucial due to the complexity of financial data and the pivotal decisions hinging on its analysis. These models signify a notable advancement in Natural Language Processing (NLP) and have attracted significant interest for their prospective utilization in the financial domain. However, before widespread adoption can occur, it is imperative to assess the efficacy and suitability of LLMs for addressing the unique challenges present within the financial domain. The financial industry demands high accuracy, reliability, and interpretability in data analysis, as decisions based on financial information can have far-reaching implications. Hence, it is imperative to assess the efficacy of LLMs in financial endeavors like sentiment analysis, trend prediction, and risk evaluation to determine their suitability for practical implementation. Additionally, assessing how well LLMs can understand and interpret financial documents, including regulatory filings, financial statements, and market reports, is essential for determining their utility in practical financial scenarios. LLMs have demonstrated promising capabilities in understanding and processing financial data. Their adeptness at swiftly and accurately processing extensive textual data renders them apt for tasks like document summarization, sentiment analysis, and extracting information from financial reports. Moreover, recent progress in LLMs, particularly in zero-shot learning and few-shot learning techniques, has broadened their scope of applications within the financial sector. One of the key considerations when evaluating the suitability of LLMs for financial tasks is their question-answering capability. In finance, question-answering involves extracting relevant information from financial documents or datasets to provide accurate responses to queries posed by users. This capability is essential for investment research, financial analysis, and risk management. The effectiveness of LLMs in financial question answering relies on various factors such as the quality of training data, question complexity, and the model's domain-specific knowledge. Assessing the performance of LLMs in question-answering tasks requires comprehensive evaluation metrics and benchmark datasets tailored to the financial domain. Despite the progress made in developing LLMs for NLP tasks, there are still challenges to overcome when applying these models in the financial domain. One such challenge is the need for domain-specific fine-tuning and customization to ensure optimal performance in financial tasks. Additionally, providing the robustness and interpretability of LLMs in financial applications is essential to build trust and confidence among users and stakeholders (Li et al., 2023)[5].

1.4 Problem Statement

In the rapidly evolving landscape of financial analysis, integrating Artificial Intelligence (AI) technologies, particularly Large Language Models (LLMs), has become increasingly prevalent. However, despite the advancements in natural language processing (NLP) capabilities exhibited by LLMs, their effectiveness in addressing domain-specific tasks, such as financial question answering, remains uncertain. The financial sector demands accurate and reliable information retrieval systems to navigate complex financial contexts and provide relevant insights. Therefore, the problem statement revolves around evaluating the suitability of LLMs, specifically Llama2-7b and Gemma-7b, in addressing financial question-answering tasks. Therefore, the central research question arises as follows: To what extent do Llama2-7b and Gemma-7b, both state-of-the-art Large Language Models, demonstrate efficacy in addressing financial question-answering tasks?

1.5 Overview

The financial sector is experiencing a transformative shift with the integration of Artificial Intelligence (AI) technologies. Large Language Models (LLMs) represent a promising avenue for innovation within this domain. These advanced AI models excel at natural language processing (NLP) tasks, including understanding complex contexts, answering inquiries, and generating human-quality text. Their potential to revolutionize financial analysis by facilitating efficient and accurate information retrieval has garnered significant interest. This research delves into the effectiveness of LLMs in answering financial question-answering tasks. The study specifically focuses on the capabilities of Llama2-7b-chat-hf (Touvron et al., 2023)[6] by Meta and Gemma 7b- it(Team Gemma et al., 2024)[7] by Google. Both models are state-of-the-art LLMs with demonstrated proficiency in general NLP tasks. However, their performance in domain-specific applications, such as financial question answering, remains more to be explored. This investigation employs a multi-pronged approach to evaluate the suitability of LLMs for answering financial questions. Three distinct evaluation methodologies are utilized: zero-shot prompt engineering, few-shot prompt engineering, and supervised fine-tuning. The performance of Llama2-7b-chat-hf and Gemma-7b-it on financial question answering is evaluated using three key metrics: ROUGE-L, cosine similarity, and human evaluation. The results of this study are anticipated to offer valuable insights into the present capabilities of LLMs in addressing financial queries. Preliminary analyses suggest that while recent LLMs exhibit promising capabilities in general NLP tasks, their accuracy and relevance in financial question-answering are limited. This underscores the necessity for continued research and development endeavors to refine LLMs' proficiency within this domain. This study adds to the ongoing investigation into utilizing LLMs in financial contexts. The results emphasize the importance of advancing LLM capabilities and crafting specialized training datasets to fully leverage their efficacy in financial question-answering tasks.

1.6 Motivation

The incorporation of Artificial Intelligence (AI) innovations is revolutionizing the financial industry, presenting transformative opportunities. Large Language Models (LLMs) are a particularly promising avenue amid this evolving landscape. These sophisticated AI models showcase impressive capabilities across a spectrum of natural language processing (NLP) tasks, encompassing nuanced contextual comprehension, precise responses to queries, and text generation akin to human quality. Their capacity to streamline financial analysis through efficient and precise information retrieval has attracted considerable attention within the industry. This research explores the efficacy of LLMs in addressing financial question-answering tasks, focusing specifically on two state-of-the-art models: Llama2-7b and Gemma-7b. While these models have demonstrated impressive performance in general NLP tasks, their suitability for domain-specific applications, such as financial question answering, remains relatively uncharted territory. Preliminary analyses suggest that while recent LLMs exhibit promising capabilities in general NLP tasks, their performance in answering financial questions may be constrained. This emphasizes the critical need for continued research and development efforts to improve the versatility and effectiveness of LLMs in this field. By actively contributing to exploring LLMs in financial contexts, this study seeks to drive progress in LLM capabilities and the development of specialized training datasets tailored to the specific requirements of financial question-answering tasks.

1.7 Objective of the Thesis

The objective of this thesis is to conduct a thorough assessment of the effectiveness and applicability of cutting-edge Large Language Models (LLMs) in addressing financial inquiries. The primary goals can be outlined as follows:

- **Assessing LLMs' Proficiency:** This study aims to assess the proficiency of prominent LLMs, specifically Llama2-7b and Gemma-7b, in accurately answering various financial questions. By subjecting these models to rigorous evaluation methodologies, including zero-shot prompt engineering, few-shot prompt engineering, and supervised fine-tuning, their aptitude for understanding and responding to financial inquiries will be thoroughly scrutinized.

- **Performance Evaluation Metrics:** Utilizing established metrics such as ROUGE-L, cosine similarity, and human evaluation, the work seeks to quantitatively measure the accuracy, relevance, and coherence of the generated responses provided by LLMs. Through meticulous analysis, the effectiveness of these models in extracting meaningful insights from financial data and delivering informative responses will be gauged.

- **Identification of Strengths and Limitations:** By systematically examining LLMs' performance across various financial question-answering scenarios, the thesis aims to identify the strengths and limitations of these models. This involves discerning their unthinkingly domain-specific terminology, navigating complex financial documents, and generating contextually appropriate responses.

- **Implications for Industry Practice:** Through disseminating research findings and recommendations, the thesis seeks to inform industry practitioners about the potential benefits and challenges of integrating LLMs into financial workflows. By highlighting the practical implications of employing LLMs for financial information retrieval, decision-making processes can be enhanced, improving efficiency and efficacy within economic organizations.

Through a rigorous evaluation of prominent LLMs, this study aims to elucidate their capabilities and constraints in addressing financial queries. By quantitatively measuring their performance and delineating their strengths and limitations, the research offers insights crucial for industry practitioners. These insights inform decision-makers about the feasibility of integrating LLMs into financial workflows, enhancing organizational efficiency and decision-making processes in the financial sector.

1.8 Challenges and Limitations

The difficulties of using a machine learning model for prediction without hyperparameter tuning can include :

- **Limited Availability of Domain-Specific Data:** The scarcity of annotated financial question-answer datasets may constrain the comprehensive evaluation of LLMs' performance in domain-specific contexts.

- **Interpretability of LLM Outputs:** The inherent complexity of LLMs poses challenges in interpreting and validating the accuracy of generated responses, potentially leading to concerns regarding model transparency and trustworthiness.

- **Variability in Financial Terminology:** The diverse terminology and nuanced language used in financial documents may pose difficulties for LLMs in accurately understanding and contextualizing queries, affecting the quality of generated answers.

- **Computation Limitations:** The computational resources required for training and fine-tuning LLMs on large-scale financial datasets may be prohibitive, posing challenges in conducting comprehensive experiments and analyses.

- **Generalization to Specific Financial Domains:** LLMs trained on general NLP tasks may struggle to generalize effectively to specific financial domains, potentially limiting their applicability and performance in real-world financial question-answering scenarios.

- **Ethical Considerations:** The ethical considerations pertaining to utilizing LLMs in financial contexts, including the risk of biased outputs, privacy infringements, and the ethical handling of confidential financial information, pose substantial hurdles. Addressing these challenges is imperative to guarantee responsible and ethical deployment of AI within the financial industry.

Despite computational challenges and the need for domain-specific adaptation, LLMs hold promise for revolutionizing information retrieval in the financial sector, provided ethical considerations are addressed, and further research is conducted to enhance their performance and applicability.

1.9 Thesis Organization

The report is organized into 6 chapters, including this chapter: ***Introduction*** where all the related topics are discussed, which are needed for understanding the research work. The outline of rest of the works are organized as follows:

Chapter 2

Topic - Background

The introduction provides an overview of Natural Language Processing (NLP) and its applications across diverse industries, highlighting the significance of Large Language Models (LLMs) like Gemma and Llama2. It explores key NLP algorithms, challenges LLMs face, and innovative solutions such as quantization and parameter-efficient fine-tuning.

Chapter 3

Topic - Literature Review

This literature review synthesizes the current state-of-the-art methodologies, challenges, and future directions in LLM-driven financial QA, shedding light on the transformative potential and critical areas for refinement in this rapidly evolving field.”

Chapter 4

Topic - Materials & Methodology

This chapter outlines methodologies for leveraging large language models (LLMs) for financial question-answering (QA), aiming to revolutionize decision-making in the financial domain by harnessing deep learning and natural language processing.

Chapter 5

Topic - Result & Performance Analysis

This chapter analyzes the performance of two question-answering models, LLama2-7b and Gemma-7b, within the financial domain, utilizing metrics such as ROUGE-L score, cosine similarity, and human evaluation.

Chapter 6

Topic - Conclusion

This chapter concludes the research work, highlighting the summary, limitations, and potential future work areas.

1.10 Conclusion

In this chapter, we were provided with an overview of the upcoming study and a glimpse into the work that lies ahead. The discussion included insights into the inspiration, objectives, and research challenges that will be further elaborated in subsequent chapters.

Chapter 2

Background

2.1 Introduction

Natural Language Processing (NLP) is a facet of artificial intelligence (AI) that empowers computers to comprehend, decipher, and produce human language akin to human interaction. It encompasses algorithms and techniques for processing textual data and facilitating tasks like language translation, sentiment analysis, and information extraction. Natural Language Processing (NLP) has diverse applications across industries, revolutionizing processes and interactions through text analysis and interpretation. [8]

- **Healthcare:** NLP classifies patient data, reliably transcribes clinical notes, and expedites documentation procedures.
- **Finance:** By applying sentiment analysis to text data, NLP helps traders make decisions by forecasting market movements.
- **Customer Service:** Chatbots with NLP capabilities instantly respond to questions, improving customer satisfaction and lightening the load on support staff.
- **E-commerce:** NLP enhances on-site search functionality, guaranteeing that visitors find pertinent products with imprecise searches.
- **Legal:** NLP speeds up document reviews by automating processes and ensuring important information is not missed.

The widespread adoption of NLP across various sectors underscores its importance in streamlining processes, enhancing user experiences, and driving efficiency through intelligent text analysis and interpretation.

2.2 Most frequently used algorithms in NLP

NLP methodologies facilitate computer comprehension and processing of human language. NLP frameworks employ diverse data preparation, feature derivation, and modeling approaches. [9, 10, 11]

- **Bag of Words (BoW):** Bag of Words (BoW) is a fundamental Natural Language Processing (NLP) technique used for text analysis and feature extraction. It represents text data as a collection of words without considering grammar or word order. Each document transforms a vector, with each dimension symbolizing a distinct word and the value denoting the frequency of that word in the document. Due to its versatility, boW finds broad application in sentiment analysis, document classification, and information retrieval endeavors.

- **Tokenization:** Tokenization involves dividing a text into individual units known as tokens, typically words or subwords, and is a fundamental step in NLP that sets the stage for further analysis. This process can occur at different levels, such as the word level, character level, or subword level, depending on the task's needs. Tokenization enables efficient text processing by supporting tasks like part-of-speech tagging, named entity recognition, and syntactic analysis.

- **Term Frequency-Inverse Document Frequency (TF-IDF):** TF-IDF, or term frequency-inverse document frequency, is a statistical metric employed to assess the significance of a term within a document compared to a broader collection of documents. It amalgamates two components: term frequency (TF), which gauges the frequency of a term's occurrence within a document, and inverse document frequency (IDF), which evaluates the uniqueness or rarity of a term across the document collection. By assigning greater weight to terms prevalent within a document but infrequent across the corpus, TF-IDF proves useful in tasks like keyword extraction, document ranking, and information retrieval.

- **Word2Vec:** Word2Vec stands as a prominent word embedding method that represents words as compact vectors within a continuous vector space. By assigning similar words to proximate vectors, it adeptly encapsulates semantic associations among words. Word2Vec models undergo training on extensive text datasets using neural networks, enabling them to discern contextual nuances and semantic meanings from neighboring words. This approach finds extensive application across various tasks, including similarity detection, named entity recognition, and sentiment analysis.

- **GloVe:** GloVe, an unsupervised learning algorithm, crafts word embeddings by leveraging global co-occurrence statistics. It creates a word-context matrix from extensive corpora and refines vector representations by minimizing disparities between word vector dot products and the logarithm of co-occurrence probabilities. GloVe embeddings encapsulate syntactic and semantic nuances and have demonstrated efficacy across various applications, including word analogy detection, document clustering, and machine translation.

- **Hidden Markov Models:** Hidden Markov Models (HMMs) are statistical frameworks tailored for sequential data modeling, rendering them apt for tasks with inherent temporal dependencies like speech recognition and part-of-speech tagging. Comprising observable and latent states, HMMs employ probabilistic transitions between states. Notably, they excel in discerning the intrinsic structure within sequential data and deducing the sequence of concealed states based on observed emissions.

- **Convolutional Neural Networks(CNN):** CNNs represent a class of deep learning architectures crafted to handle structured grid-like data, including images and text. Within Natural Language Processing (NLP), CNNs find application in tasks such as text classification and sentiment analysis. Leveraging convolutional layers, they extract localized features from input sequences, complemented by pooling layers for dimensionality reduction and feature extraction. Renowned for their ability to capture spatial hierarchies in data, CNNs consistently demonstrate remarkable performance across a spectrum of NLP tasks.

- **Recurrent Neural Networks (RNN):** RNNs are specialized architectures tailored to handle sequential data by preserving internal state or memory. RNNs update their hidden state at each step, operating on input sequences incrementally, making them ideal for tasks demanding sequential modeling like language modeling, machine translation, and speech recognition. Despite their efficacy, conventional RNNs encounter challenges like vanishing or exploding gradients, hindering long-range dependency learning. Nevertheless, advancements like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures have emerged to address these limitations and enhance performance across NLP tasks.

- **Autoencoders:** Autoencoders are neural network structures devised for unsupervised learning, focusing on deriving efficient data codings. They aim to acquire a condensed representation of input data, often utilized for tasks like dimensionality reduction or feature extraction. These architectures comprise two main components: an encoder network responsible for compressing input data into a latent space representation, and a decoder network tasked with reconstructing the input data from the latent space representation. This process allows autoencoders to learn compact and meaningful representations of input data without requiring labeled supervision.

- **Transformers:** Transformers represent a category of deep learning model architectures engineered to process sequential data by capturing extensive dependencies. They leverage self-attention mechanisms to assess the significance of various input components in generating output representations. Renowned for their proficiency in capturing contextual nuances, transformers have emerged as pivotal tools in NLP endeavors. Their adeptness enables a spectrum of tasks including language translation, text generation, and sentiment analysis, marking a significant advancement in the field of natural language processing.

In essence, NLP has revolutionized our ability to analyze and interpret textual data. With advanced techniques like Hidden Markov Models and RNN, NLP drives innovation in various fields, facilitating deeper understanding and utilization of human language.

2.3 Large Language Models (LLM)

Large language models (LLMs), abbreviated as LLMs, denote machine learning models trained on extensive text datasets. This training equips them to produce text that resembles human language, execute diverse language-related tasks proficiently, and attain a deeper comprehension of human language nuances. LLMs find applications across a spectrum of scenarios. [12]

- **Text generation:** A Large Language Model (LLM) trained on diverse topics can generate coherent text on any subject it has been exposed to during training.
- **Translation:** LLMs adept in multiple languages can accurately and fluently translate text from one language to another.
- **Content summary:** LLMs excel at condensing lengthy blocks of text or multiple pages into concise and informative summaries.
- **Classification and categorization:** LLMs can accurately classify and categorize content based on predefined criteria, aiding in organizing and managing large volumes of data.
- **Sentiment analysis:** LLMs can conduct sentiment analysis to interpret a text's emotional tone and sentiment, enabling a more nuanced comprehension of user intent or content reaction.

Table 2.1: Best Large Language Models in 2024 [13]

LLM	Developer	Popular apps that use it	# of parameters	Access
GPT	OpenAI	Microsoft, Duolingo, Stripe, Zapier, Dropbox, ChatGPT	175 billion	API
Gemini	Google	Some queries on Bard & Nano: 1.8	3.25 billion; others unknown	API
PaLM 2	Google	Google Bard, Docs, Gmail, and other Google apps	340 billion	API
Llama 2	Meta	Undisclosed	7, 13, and 70 billion	Open source
Claude 2	Anthropic	Slack, Notion, Zoom	Unknown	API
Falcon	Technology Innovation Institute	Undisclosed	1.3, 7.5, 40, and 180 billion	Open source
MPT	Mosaic	Undisclosed	7 and 30 billion	Open source
Mixtral 8x7B	Mistral AI	Undisclosed	46.7 billion	Open source

LLMs are powerful tools adept at generating text, translating languages, summarizing content, categorizing data, and analyzing sentiment. Their versatility and accuracy make them indispensable in various applications.

2.4 Llama2-7b and Gemma-7b

Large Language Models (LLMs) have risen as a potent tool for tasks in natural language processing. This study delves into two prominent LLMs: Gemma, developed by Google AI, and Llama 2, released by Meta. We aim to understand their strengths and weaknesses through a comparative analysis.

- **Capabilities in different domain:** Llama2 and Mistral Gemma's language understanding and generation performance across various capabilities are compared to similarly sized open models (7B) on standard academic benchmark evaluations.

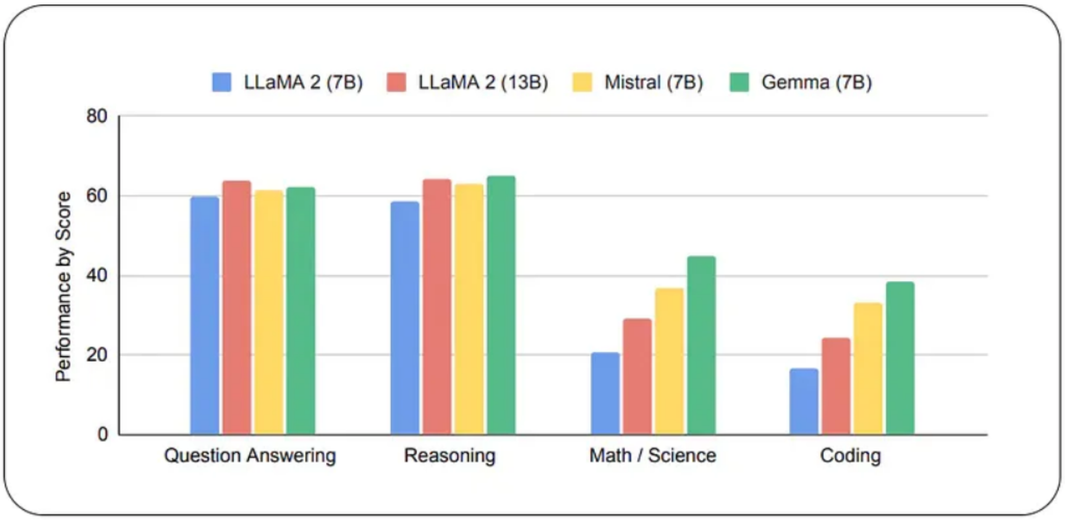


Figure 2.1: Gemma vs. Llama vs. Mistral: A Comparative Analysis [14]

- **Architectural Innovations and Training Methodologies:** Gemma distinguishes itself through innovative enhancements such as multi-query and multi-head attention mechanisms, RoPE embeddings, GeGLU activations, and strategic normalizer placement. These architectural advancements equip Gemma to tackle complex tasks across extensive context lengths efficiently. Gemma refines its instruction-tuned models through supervised fine-tuning and reinforcement learning from human feedback, leveraging a vast training dataset of 6 trillion tokens, primarily consisting of English language texts. Conversely, Llama2, developed by Meta, prioritizes scalability and efficiency with its transformer-based architecture. Engineered for adaptability across diverse tasks with minimal fine-tuning requirements, Llama2 models are trained on a comprehensive dataset, emphasizing general applicability and seamless integration into various applications.

- **Benchmark Accuracy and Task Proficiency:** Research suggests that Gemma exhibits superior accuracy compared to Llama 2 on various benchmarks. Notably, Gemma's 7-billion parameter model achieves a general accuracy of 64.3%, surpassing Llama 2 in tasks requiring reasoning, mathematical problem-solving, and other categories. This indicates that Gemma might be better suited for applications demanding a deeper understanding of the subject matter.

Table 2.2: Performance of Gemma on Various Benchmarks[15]

Capability	**Benchmark**	**Description**	**Gemma(7B)**	**Llama-2(7B)**	**Llama-2(13B)**
General	MMLU 5-shot, top-1	Representation of questions	64.3	45.3	54.8
Reasoning	BBH	Multi-step reasoning tasks	55.1	32.6	39.4
	HellaSwag 0-shot	Commonsense reasoning	81.2	77.2	80.7
Math	GSM8K	Basic arithmetic	46.4	14.6	28.7
	MATH 4-shot	Challenging math problems	24.3	2.5	3.9
Code	HumanEval pass@1	Python code generation	32.3	12.8	18.3

- **Model Configuration:** LLama2 and Gemma exhibit contrasting model configurations, each tailored to address specific computational and task-related requirements.

Table 2.3: Comparison of LlamaConfig and GemmaConfig[16]

Configuration	**LlamaConfig**	**GemmaConfig**
Vocabulary Size	32000	256000
Hidden Size	4096	3072
Intermediate Size	11008	24576
Number of Hidden Layers	32	28
Number of Attention Heads	32	16
Number of Key-Value Heads	None	16
Head Dimension	-	256
Hidden Activation	silu	gelu_pytorch_tanh
Max Position Embeddings	2048	8192

2.5 Quantization of Large Language Models

Large Language Models (LLMs) face challenges when integrated into real-world applications due to their inherent complexities, largely stemming from their size and computational demands. Despite excelling in tasks like text generation and translation, the extensive resource requirements of LLMs constrain their accessibility and deployment across diverse platforms. Recent research endeavors have explored the potential of quantization techniques to alleviate these challenges and democratize access to LLMs. Quantization offers a promising avenue to optimize LLMs' efficiency and resource utilization. By reducing the numerical precision of model parameters from the standard 32-bit floating-point format (FP32) to lower bit-width representations like 8-bit (INT8) or 4-bit (INT4), quantization significantly reduces model size, leading to several key benefits. Firstly, quantization results in a reduced memory footprint, enabling deployment on devices with limited memory resources, thereby expanding the reach of LLMs to edge devices and mobile platforms. This reduction in memory usage enhances computational efficiency, resulting in faster inference speed and improved user experience. Additionally, quantization democratizes access to LLMs by minimizing hardware requirements, empowering a wider range of users to leverage these advanced models for various applications. However, quantization introduces a trade-off between resource savings and model accuracy. Lower precision representations inevitably introduce quantization errors, which may impact model performance. The severity of this impact depends on factors such as model architecture, chosen quantization method, and specific task requirements. Researchers continuously strive to develop robust quantization techniques that minimize performance degradation while maximizing resource savings.[17] Focusing specifically on 4-bit quantization represents a significant advancement in this pursuit. Integrating 4-bit quantization into LLMs promises even further benefits than higher-precision quantization methods. 4-bit representations offer an ultra-low memory footprint, enabling deployment on more resource-constrained devices. Additionally, the reduced computational burden associated with 4-bit quantization leads to faster inference speeds, enhancing the overall efficiency of LLM applications. Despite these advantages, achieving effective 4-bit quantization for LLMs remains challenging. Strategies such as adaptive quantization, post-training quantization, and quantization-aware training are being explored to minimize performance degradation while maximizing compression ratio.[18]

The loading process of models integrated with 4-bit quantization parameters follows specific configurations. Utilizing the BitsAndBytesConfig module[19], the load_in_4bit parameter is set to True, indicating the adoption of 4-bit precision. This choice ensures the model operates with reduced numerical precision, enhancing computational efficiency and memory utilization. The bnb_4bit_compute_dtype parameter is configured based on the specified compute_dtype variable, dictating the data type utilized for computational operations during model loading and subsequent processing. These configuration settings ensure the effective integration of 4-bit quantization techniques into the model loading process, facilitating efficient resource utilization and model performance.

2.6 Supervised Fine Tuning

Large language models (LLMs) have profoundly reshaped the realm of natural language processing (NLP) by harnessing extensive text data to acquire sophisticated language representations. Yet, their effectiveness in particular downstream tasks frequently demands fine-tuning on datasets tailored to those tasks. Conventional fine-tuning approaches involve adjusting all model parameters, which brings about significant computational expenses and the potential risk of catastrophic forgetting, where the model erases previously learned knowledge. Parameter-efficient fine-tuning (PEFT) (Ding et al., 2023)[20] is compared with pre-training and conventional fine-tuning methods, delineating its advantages and specific techniques. Pre-training is the foundational step, wherein LLMs are trained on extensive general-purpose text data to develop a generic language representation. While pre-training provides a solid starting point, its applicability to specific downstream tasks may be suboptimal. Traditional fine-tuning involves adjusting all parameters of the pre-trained model using the dataset specific to the downstream task. Although this approach leverages pre-trained knowledge, it poses challenges such as high computational costs and catastrophic forgetting. PEFT offers a more efficient alternative by focusing on adapting a smaller subset of model parameters relevant to the specific task. This results in reduced computational costs mitigated catastrophic forgetting, and improved generalizability compared to conventional fine-tuning.

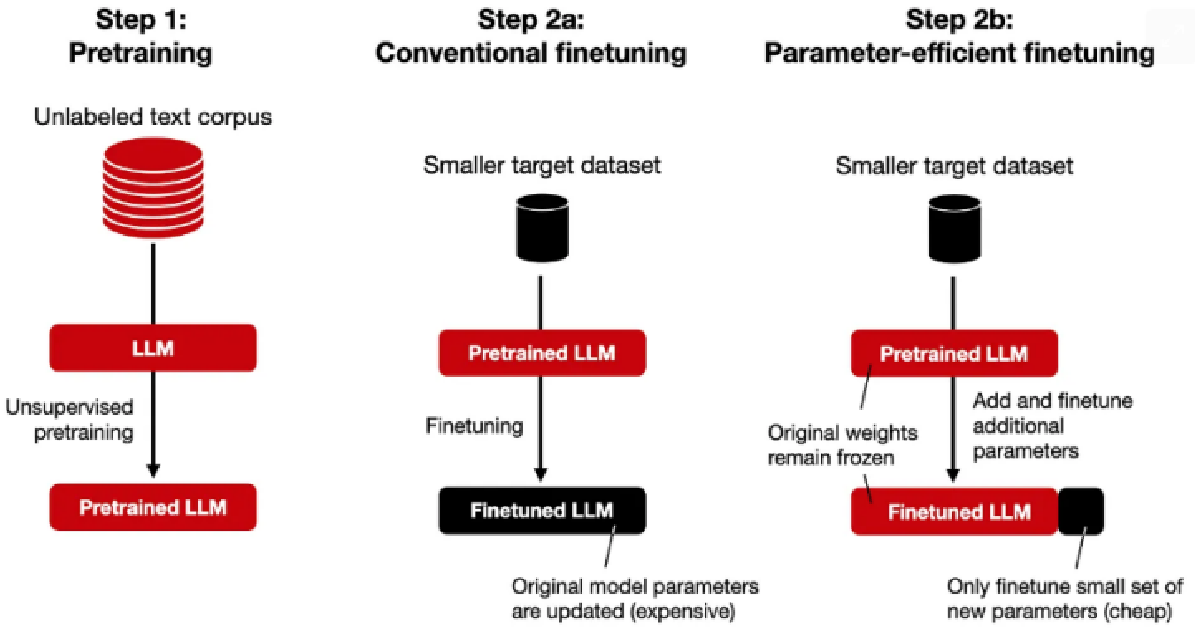


Figure 2.2: Pre-training, conventional fine-tuning and parameter efficient fine tuning[20]

Parameter-efficient fine-tuning (PEFT) not only addresses the limitations of conventional fine-tuning but also offers unique benefits for specific applications.PEFT's diminished computational demands facilitate the deployment of large language models (LLMs) in resource-constrained settings, such as mobile phones or edge computing platforms, making it a viable option for such environments. It facilitates transfer learning across related tasks by selectively adapting a subset of parameters while preserving pre-trained knowledge, which is particularly advantageous in situations with limited downstream task data.

Low-rank adapters (LoRA) (Hu et al., 2021)[21] represent a parameter-efficient fine-tuning (PEFT) method that integrates low-rank adapter modules into the process of adapting large language models (LLMs) to particular downstream tasks. This approach aims to reduce computational costs while maintaining competitive performance levels. The mechanism of LoRA involves the introduction of low-rank adapter modules alongside the pre-trained LLM. These adapter modules are designed to have significantly fewer parameters compared to standard adapter modules. By employing low-rank parameterization techniques, LoRA achieves parameter efficiency by representing adapter weights in a low-dimensional space. Low-rank parameterization involves decomposing the weight matrices of adapter modules into low-rank matrices, effectively reducing the number of parameters required for adaptation. This compression technique produces a more compact representation of adapter weights while retaining essential task-specific information. An essential benefit of LoRA lies in its capacity to alleviate the computational overhead linked with fine-tuning extensive LLMs. LoRA markedly diminishes the volume of parameters necessitating updates during fine-tuning, resulting in expedited training durations and decreased resource demands through employing low-rank parameterization.

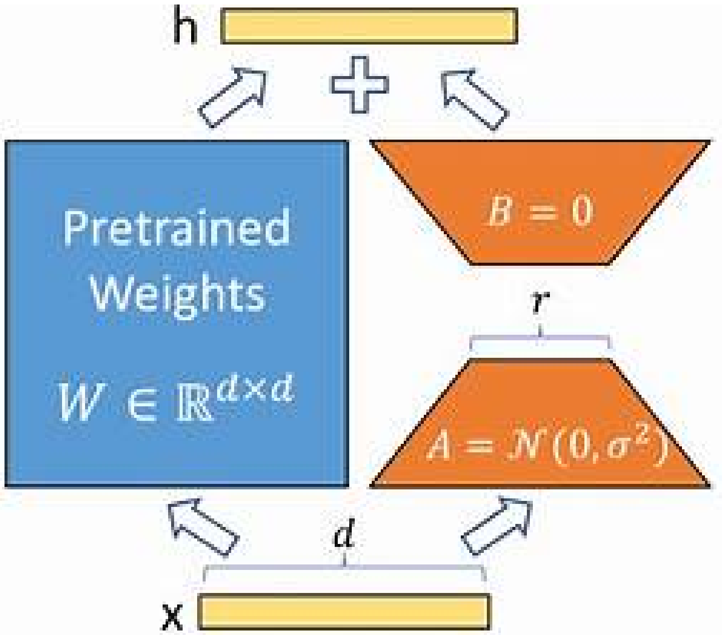

h
Pretrained
Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

Figure 2.3: Low-rank Adapters[21]

Despite the parameter reduction, LoRA maintains competitive performance levels compared to standard adapter modules. This is achieved by carefully selecting the low-rank dimensions and optimizing the adapter weights to capture task-specific information effectively. The reduced computational overhead allows LLMs equipped with LoRA adapters to be deployed more efficiently without compromising performance.

2.7 Challenges and Limitations of LLMs

Large Language Models (LLMs) have exhibited notable progress across diverse domains, encompassing text generation, translation, and content summarization. However, despite their significant progress, LLMs encounter challenges and limitations that hinder their effectiveness and applicability in real-world scenarios. [22]

- **Technical Limitations:** LLMs often confront technical limitations that affect their accuracy and contextual understanding. One such challenge is domain mismatch, where models trained on broad datasets struggle to comprehend specific or niche subjects due to inadequate data representation. Consequently, this limitation leads to inaccuracies or generic responses when addressing specialized knowledge domains. Moreover, LLMs encounter difficulties predicting less common words or phrases, impacting their text generation capabilities and performance in translation and technical documentation tasks.

- **Real-time Translation Efficiency:** While LLMs have made significant strides in translation accuracy, their computational demands pose challenges for real-time translation, particularly for languages with complex grammatical structures or limited training data. The computational resources required to process and generate real-time translations can strain system capabilities and hinder efficiency.

- **Hallucinations and Bias:** LLMs occasionally exhibit ”hallucination” behavior, generating erroneous or fictitious information. This phenomenon poses risks, as evidenced by incidents like Air Canada’s chatbot erroneously informing customers about a non-existent refund policy. Furthermore, Large Language Models (LLMs) might unintentionally reinforce and magnify biases present in their training data, leading to discriminatory or offensive outputs.

- **Scalability and Environmental Impact:** The scalability of LLMs raises concerns regarding their environmental impact, particularly in terms of energy consumption. Training large language models requires substantial energy resources, with systems like GPT-3 consuming significant megawatthours of energy. Moreover, operating LLMs on a large scale further exacerbates energy consumption, posing sustainability challenges and contributing to environmental concerns.

While LLMs offer immense potential for advancing natural language processing tasks, addressing these challenges and limitations is crucial to enhance their efficiency, accuracy, and ethical implications in real-world applications. Future research efforts should focus on mitigating technical constraints, reducing environmental impact, and promoting fairness and inclusivity in LLM development and deployment.

2.8 Conclusion

The introduction delineates the essential role of Natural Language Processing (NLP) in contemporary AI applications spanning various industries. It underscores the prominent algorithms in NLP and the significance of Large Language Models (LLMs), particularly Gemma and Llama2, due to their architectural innovations, benchmark achievements, and model configurations. Despite their progress, challenges such as technical constraints, efficiency in real-time translation, biases, and environmental implications underscore the imperative for ongoing research to tackle these issues and fully harness the potential of LLMs in practical applications.

Chapter 3

Literature Review

3.1 Introduction

This section presented herein offers a comprehensive examination of recent advancements and challenges in financial question answering (QA) utilizing large language models (LLMs). This review synthesizes findings from diverse studies, from the introduction of transformative LLM architectures to their application in addressing nuanced financial inquiries. By scrutinizing a spectrum of methodologies and outcomes, this review elucidates the current landscape of LLM-driven financial QA while pinpointing critical areas for further investigation and refinement. By thoroughly examining academic contributions, this review seeks to offer a nuanced comprehension of the capabilities, limitations, and future directions of LLM technology in facilitating accurate and reliable financial analyses.

3.2 Related Works

The transformer represents a neural network structure initially presented in the paper titled ”Attention Is All You Need” (Vaswani et al., 2023) [23]). It has proven to be a more reliable and efficient technique when compared to earlier methods like Word2Vec and Recurrent Neural Networks (RNNs). The transformer was designed to simulate neural network architecture using two effective methods - encoder-only or encoder-decoder designs. The most basic attention mechanism, self-attention, establishes correlations based on positional significance within a sequence. Utilizing the multi-head self-attention mechanism, the transformer architecture effectively handles long-range dependencies.

Among the noteworthy transformer models, GPT (Generative Pre-trained Transformer), BERT (Bidirectional Encoder Representations from Transformers), and T5 (Text-to-Text Transfer Transformer) (Raffel et al., 2020)[24] have attracted considerable attention due to their exceptional performance in NLP tasks via transfer learning[25]. BERT, an encoder-only framework introduced by (Devlin et al., 2018)[26], excels in generating contextually sensitive token embeddings derived from extensive pre-training on corpora such as Wikipedia and Google's BooksCorpus (Bandy et al., 2021)[27]. Conversely, BART (Bidirectional and Auto-Regressive Transformers), proposed by (Lewis et al.,2019)[28], encompasses both encoder and decoder components, enabling sequence-to-sequence modeling and text generation. BART's distinctive dual-stage training, involving text perturbation and subsequent reconstruction, enhances its capacity for intricate transformations and text generation. The GPT series (Radford et al., 2018)[29], including models like GPT-3 and GPT-4, are decoder-centric architectures pre-trained for predictive tasks involving subsequent tokens. Noteworthy is GPT -3's staggering 175 billion parameters, surpassing BERT and BART in scale, while GPT-4 scales even further, boasting 70 trillion parameters. These models utilize self-supervised learning and Reinforcement Learning from Human Feedback (RLHF) to enhance outputs, leading to fluency and responses akin to those generated by humans. Furthermore, alongside structural differences, models such as GPT-3 and GPT-4 demonstrate remarkable levels of generalization, facilitating zero-shot and few-shot learning. This differs from the capabilities of BERT and BART, which excel in fine-tuning for specific tasks using annotated data.

Question Answering (QA) is a subfield of Machine Reading Comprehension (MRC) that focuses on not only understanding a question but also generating an answer based on factual information within a provided passage (Joshi et al., 2017)[30]. Two main approaches exist for building QA systems: extractive and generative. Extractive models identify the answer passage directly within the context document. This approach is efficient for factual ”Right There” questions where the answer exists within the text. However, generating large labeled datasets for training is a challenge. Generative or abstractive models create a natural language answer that addresses the question but may not directly match a specific passage. This allows for more human-like responses. Recent advancements in generative pre-trained transformers have opened exciting possibilities for QA systems. Still, challenges include potential biases and factual inconsistencies (”hallucinations”) in the generated responses (Rajpurkar et al., 2016)[31]. Open Domain Question Answering (ODQA) focuses on finding answers to factual questions within vast collections of unstructured documents. This is particularly relevant for knowledge-intensive professions like medicine, law, and academia (Kosti et al., 2021)[32]. ODQA systems can leverage various data modalities, including natural language documents, structured data, and semi-structured data. Open-book QA allows the system to access an external corpus of documents during inference, while closed-book QA does not. Information retrieval (IR) techniques are often employed to identify relevant passages within the document collection. Zero-Shot Open Book QA presents an additional challenge, requiring the system to answer questions without supervised machine learning on the specific dataset. This focus on generalizability and flexibility pushes the boundaries of current QA technology (Zhang et al., 2022)[33].

The utilization of Large Language Models (LLMs) in the financial industry has experienced notable acceleration in recent years. A notable example is the release of BloombergGPT in early 2023 (Wu et al., 2023)[34]. This LLM garnered considerable attention for its superior performance on benchmarks relevant to the financial domain. BloombergGPT outperforms other LLMs in tasks related to general reasoning, benchmarks specific to the financial sector, and even in handling proprietary datasets owned by Bloomberg. This achievement is attributed to its massive architecture, boasting 50 billion parameters Furthermore, its training regimen leverages a unique combination of data sources, incorporating 363 million tokens from industry-specific datasets curated by Bloomberg alongside 345 million tokens from general natural language datasets. This targeted training approach equips BloombergGPT with a robust understanding of both general language and the intricacies of financial terminology and concepts. FinGPT, developed by (Yang et al. in 2023)[35], is a financial language model that incorporates information from diverse sources like social media, financial filings, trends, and academic datasets to offer insights for financial analysis. It emphasizes real-time data processing, fine-tuning of LLMs, and applications in robo-advising, quantitative trading, sentiment analysis, and more. The model aims to democratize access to advanced financial modeling techniques and personalized financial advice by promoting open-source values, low-cost adaptation, and access to high-quality financial data. The ConFIRM framework (Choi et al., 2023)[36], tailored for conversational financial information retrieval, harnesses the power of large language models to classify query intent and label knowledge bases in finance. Demonstrating remarkable accuracy rates surpassing 90%, it presents an efficient solution for extracting precise query intent with minimal data requirements. By curating a dataset of finance-specific question-answer pairs, ConFIRM emerges as a pragmatic tool for regulatory compliance in finance, reducing the need for extensive human intervention. Shah et al., (2022) [37] introduced novel financial language models, FLANG-BERT and FLANG-ELECTRA, specifically tailored to financial vocabulary to elevate their efficacy in financial language-related tasks. These models perform superior to their predecessors across sentiment analysis, classification, and question-answering domains. FLANG incorporates financial-specific keywords and phrases to optimize token masking and introduces the Financial Language Understanding Evaluation (FLUE) benchmarks for assessing model proficiency in financial contexts. Pre-trained on a combination of general English and finance-specific datasets, the FLANG model demonstrates improved performance across a range of financial natural language processing (NLP) tasks. The TAGOP model, introduced by (Zhu et al., 2021)[38], showcases notable proficiency in addressing queries within the TAT-QA dataset, renowned for its hybrid nature encompassing both tabular and textual contexts within finance. TAGOP surpasses baseline models by integrating tabular and textual data, exhibiting superior Exact Match and F1 scores. The model excels in finance-related numerical reasoning tasks by leveraging sequence tagging and aggregation operators.

Financial question answering (QA) benchmarks play a pivotal role in evaluating abilities to interpret financial data accurately, focusing on aspects like sentiment analysis and opinionated QA. While FiQA (Maia et al., 2018)[39] initially aimed to assess model performance in interpreting financial data, its emphasis on sentiment analysis limits its scope, as financial inquiries encompass a broader spectrum of topics concerning companies. In response to this limitation, FinQA (Chen et al., 2021)[40] emerged as a high-caliber open-access dataset containing more than 8,000 question-answer pairs curated by financial professionals, enriching the spectrum of evaluations in financial question-answering. Subsequently, Chen et al. (2022)[41] expanded upon FinQA with ConvFinQA, introduced in 2022, which introduced a more realistic and intricate testing setup by incorporating interactions comprising multiple questions dependent on prior exchanges. This advancement, comprising 3,892 conversations and 14,115 questions, reflects a nuanced approach to evaluating models' aptitude to handle complex and sequential financial inquiries. Alvarado et al. (2015)[42] presented a dataset designed specifically for named entity recognition of credit risk attributes in financial documents, fulfilling the requirement for specialized datasets for domain-specific tasks. This dataset is valuable for evaluating models' abilities to identify critical information within financial documents, contributing to enhanced risk management and decision-making processes. Moreover, Callanan et al. (2023) [43] adopted a novel approach by assessing the capability of large language models (LLMs) to respond to simulated exam queries for the Chartered Financial Analyst (CFA) Program, levels I and II. Despite the lack of publicly accessible passing standards for the CFA, the study's results indicate that leading LLM implementations show significant promise in attaining satisfactory proficiency levels in these demanding assessments, highlighting the adaptability of LLMs in tackling domain-specific hurdles beyond conventional question-answering tasks.

Kamalloo et al. (2023)[44] examined the limitations of lexical matching in accurately assessing model performance, especially regarding the nuanced responses generated by LLMs. The authors advocated human evaluation as a more reliable method for precise assessments, highlighting the deficiencies of automated evaluation techniques like lexical matching. They proposed a zero-shot prompting approach as an alternative to human evaluation, aiming to address the limitations of lexical matching. They stressed the need for more robust evaluation methodologies in open-domain question answering, particularly due to challenges posed by syntactic response variations. Their investigation involved evaluating various models, including BERT-based models and LLMs, revealing discrepancies in model performance depending on the evaluation method employed. Zhang et al. (2021) [45] presented a methodology for fine-tuning and assessing large language models (LLMs) customized for specialized monetization objectives. Their approach entailed a hybrid fine-tuning strategy that integrated in-domain and general-purpose data to balance linguistic proficiency and domain-specific knowledge. The researchers constructed a thorough evaluation framework consisting of 45 tailored questions to evaluate performance across various aspects including reliability, coherence, and business applicability. Furthermore, they investigated the impact of model size and ongoing training on performance metrics to guide effective resource allocation during the fine-tuning phase. They found that Llama-2 generally exhibited superior reasoning abilities to Vicuna models, particularly regarding accuracy scores. However, Vicuna models, notably Vicuna-v1.3, demonstrated notable strengths in courtesy and safety criteria. Zhang et al. (2021)[46] explored the capabilities of LLaMA-2 in executing table-based fact verification (TFV) tasks across various scenarios, including zero-shot, few-shot, and instruction-tuning settings. The study assessed LLaMA-2's efficacy in directly engaging in TFV tasks and compared its performance against other LLMs and program-based approaches. Results indicated that while LLaMA-2, boasting 7B parameters, exhibited enhanced accuracy post-fine-tuning, it still trailed certain smaller-scale table-based Probabilistic Logic Models (PLMs). Notably, instruction fine-tuning was identified as a potential avenue for augmenting LLaMA-2's performance, albeit challenges such as hallucination for complex queries were noted. Alawwd et al. (2024) [47] investigated the efficacy of large language models (LLMs) in the Textbook Question Answering (TQA) task utilizing fine-tuning and retrieval augmented generation (RAG) techniques as proposed by Lewis et al. (2020) [48]. It involves fine-tuning the Llama-2 model using domain-specific data from the CK12-QA dataset to enhance its reasoning prowess and accuracy in addressing TQA queries. Additionally, the incorporation of RAG strategies aims to refine the contextual output generated by the LLM, mitigating the ”out-of-domain” challenge encountered in TQA. The study underscores the significance of harnessing LLM capabilities while acknowledging the imperative of fine-tuning domain-specific datasets for optimal performance. In their recent study, Adewumi et al. (2024)[49] directed their focus toward assessing the efficacy of three contemporary Large Language Models (LLMs) – LLaMA-2-13B, Mixtral 8x7B, and Gemma-7B – in author attribution endeavors concerning brief excerpts extracted from renowned literary works. The investigation specifically delved into the performance evaluation of LLaMA-2-13B and Gemma-7B across various authors. Notably, LLaMA-2-13B demonstrated fluctuating performance levels across different authors, with observed correlations between accuracy and the newly introduced Simple Hallucination Index (SHI). Conversely, Gemma-7B presented mixed outcomes concerning both accuracy and SHI, suggesting potential operational efficacy limitations. A significant contribution of the study was the introduction of the SHI metric, which served as a systematic tool for gauging instances of false attribution, thus underscoring the imperative of mitigating hallucination phenomena within LLMs to bolster their reliability within author attribution tasks.

The Finance Bench dataset, as introduced by Islam et al. (2023) [50], stands as a comprehensive repository tailored for financial question-answering tasks, housing a total of 10,231 cases. Recognized for its invaluable contribution to financial inquiry, this dataset underwent meticulous evaluation, leveraging a spectrum of large language models (LLMs), among which Llama2 emerged as a prominent contender. Extensive assessments were conducted to ascertain the efficacy of Llama2 in financial QA tasks across diverse configurations and prompt sequences. The complexities inherent in the Finance Bench dataset pose significant challenges for LLMs. Certain questions, though ecologically valid, tend to be overly simplistic, thus inflating model performance metrics within the benchmark framework. Moreover, the pervasive ambiguity surrounding correct responses further exacerbates the challenge, casting doubt on the reliability of model-generated outputs, particularly in the realm of financial question answering. These complexities underscore the need for robust evaluation methodologies to gauge LLM performance accurately within such intricate contexts. Despite advancements in LLM technology, limitations persist, particularly in models within the domain of financial QA tasks. The prevalence of erroneous responses and refusals from Llama2 and analogous models underscores inherent impediments afflicting their information retrieval and reasoning faculties, both critical for delivering accurate financial analyses. Additionally, the risk of hallucinations and the propagation of logically flawed outputs serve as poignant reminders of the formidable challenges hindering the practical deployment of LLMs within financial settings. The research conducted by Zhang et al. (2024) [51] delve into the domain of financial question answering, with a particular focus on enhancing the performance of large language models (LLMs) through few-shot learning, and fine-tuning methodologies. The study's methodology revolves around the training and evaluating of several prominent models, including GPT-3.5 Turbo, GPT4All, Llama2, and Claude, utilizing the FinanceBench dataset. A notable aspect of the study is the significant attention given to fine-tuning processes aimed at improving the accuracy of LLMs in financial question-answering tasks. Through supervised fine-tuning on the zero-shot model for GPT-3.5 Turbo and implementing promoting techniques on GPT4All and Llama2, the models are guided to assimilate specific response patterns based on example prompts, questions, evidence text, and answers. The iterative refinement facilitated by fine-tuning and promoting aims to bolster the models' performance on the FinanceBench dataset. The study highlights the importance of training models with a limited number of labeled samples in areas where they encounter challenges, resulting in notable improvements in accuracy over time. Despite the valuable insights provided by the study, several limitations warrant consideration. Firstly, the evaluation metrics employed lack detailed elaboration, potentially introducing ambiguity in assessing model performance. The lack of human evaluation presents a constraint, as human assessment is essential for gauging the caliber and applicability of model outputs, particularly within intricate domains such as financial question answering. Moreover, an insufficient explanation of the prompt development and workflow process hinders understanding how the models were trained and fine-tuned. Furthermore, the study's employing varying methods for different models may introduce inconsistencies in the evaluation process, hindering direct comparisons between models. This lack of uniformity in experimental setups may yield biased results and impede the reproducibility of the study's outcomes. Integrating human evaluation into the fine-tuning process can provide invaluable insights into the caliber and pertinence of model outputs, thereby ensuring the generation of precise and trustworthy responses. Standardizing prompt creation and fine-tuning procedures across different models can promote consistency and facilitate direct comparisons. Furthermore, exploring diverse model parameters, embedding models, and fine-tuning strategies customized for the finance-centric domain can elevate the accuracy and effectiveness of large language models in financial question-answering endeavors.

3.3 Conclusion

The literature review furnishes a thorough examination of the progressions and hurdles encountered within the realm of financial question answering (QA) employing large language models (LLMs). While recent studies showcase promising results in refining LLM performance through fine-tuning and promoting methodologies, critical limitations persist. Particularly, the lack of clarity in prompt development and experimental workflows, coupled with varying evaluation methods across studies, hinders the reproducibility and comparability of findings. Addressing these limitations is imperative for advancing the reliability and applicability of LLMs in financial QA tasks. The study prioritizes standardizing experimental procedures, incorporating human evaluation into zero-shot prompt engineering, few-shot prompt engineering & supervised fine-tuning processes, and exploring domain-specific fine-tuning strategies to enhance LLM accuracy and consistency. By confronting these challenges, the discipline can unleash the complete potential of LLMs for precise and dependable financial analysis and decision-making.

Chapter 4

Methodology & Implementation

4.1 Introduction

The study endeavors to fulfill a crucial requirement for resilient and effective financial question-answering (QA) solutions by harnessing the capabilities inherent in large language models (LLMs) like Llama2 and Gemma. In today's rapidly evolving financial landscape, extracting actionable insights from vast amounts of unstructured financial data is paramount for informed decision-making and risk management. Conventional financial analysis methods frequently encounter challenges in handling the extensive volume and intricate nature of available data, thus compelling the adoption of more sophisticated and adaptable approaches. By harnessing the power of deep learning and natural language processing, LLMs offer the potential to revolutionize financial QA by providing accurate, context-aware responses to complex queries in real-time. However, realizing this potential requires overcoming several challenges, including the need for tailored adaptation to the intricacies of financial language and concepts. The study aims to address this disparity by methodically assessing the efficacy of LLMs in answering financial questions through a multi-phase methodology. Zero-shot learning enables exploring the models' inherent capability to extrapolate knowledge to unfamiliar tasks. Conversely, few-shot learning mimics situations where only a limited number of task-specific examples are available for training. Through fine-tuning with a parameter-efficient approach, the models undergo iterative refinement to enhance performance on financial QA tasks, maximizing their utility in real-world applications. Through deepening our comprehension of the strengths and weaknesses of LLMs in financial question answering, this study endeavors to foster the development of more effective and dependable solutions for financial analysis and decision-making support.

4.2 Workflow of the research

The research begins by loading the Llama2-7b-chat-hf and Gemma-7b-it models in a 4-bit quantization format and acquiring the finance-bench dataset. The dataset forms the cornerstone for subsequent training and evaluation endeavors. First, the models undergo a zero-shot learning phase. In this phase, the system is prompted with questions and context, and tasked with generating answers on the test set without prior task-specific training. The research progresses to a few(3)–shot learning phases following zero-shot learning. Here, the system is provided with 3 question-context-answer triplets for training, simulating scenarios where only limited task-specific examples are available. The models' ability to generate answers on the test set is then evaluated. Next, the focus shifts to fine-tuning the models. Parameter-efficient fine-tuning is employed using a Lora configuration. The models undergo fine-tuning for 5 epochs, allowing for iterative adjustments of model parameters to enhance performance. The results from zero-shot, few-shot, and fine-tuning phases are then analyzed and interpreted. This entails assessing the effectiveness of each learning paradigm and evaluating how the duration of fine-tuning influences model performance.

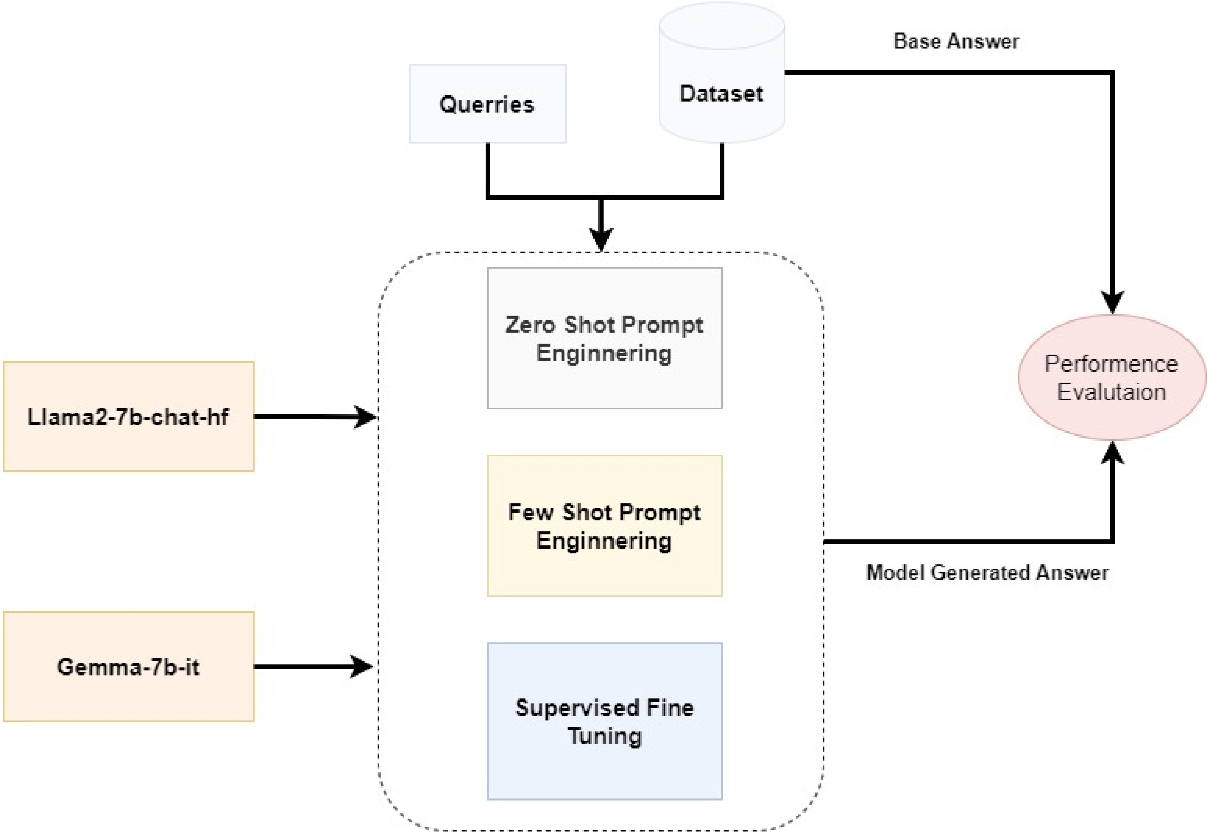


Figure 4.1: Workflow of the research

4.3 Dataset Description

The Finance Bench dataset serves as a standardized benchmark dataset specifically designed to evaluate the effectiveness of Large Language Models (LLMs) in addressing queries within the domain of financial question answering. Comprising a total of 10,231 instances, each entry within the dataset is characterized by 16 distinct columns of information. Structured to encompass a wide spectrum of financial scenarios, the dataset predominantly focuses on publicly traded companies within the United States. Central to the dataset are pivotal columns about the companies, furnishing critical details such as company names, ticker symbols, and industry sector classifications. These attributes facilitate the identification and categorization of queries based on the specific entities involved, thereby enhancing the granularity of analysis. In addition to company-specific information, the dataset incorporates columns delineating the nature of questions posed, encompassing diverse categories such as domain-specific, novel, and metric-generated inquiries. This categorization framework enables a comprehensive assessment of the varied queries prevalent within the financial domain. The dataset includes columns dedicated to reasoning types associated with each query, providing insights into the underlying logic or rationale for addressing specific inquiries. Such attributes offer a nuanced understanding of the cognitive processes in answering financial questions, enriching the dataset with valuable interpretive dimensions. Including evidence text columns within the dataset provides corroborative documentation or contextual background relevant to each query. These textual elements are pivotal in facilitating accurate response generation by LLMs, empowering them with requisite contextual information for informed decision-making. Additionally, the dataset incorporates columns detailing key financial metrics such as market capitalization, Global Industry Classification Standard (GICS) sector classification, and inclusion in the S&P 500 index for the companies under scrutiny. These attributes contribute valuable contextual insights regarding the companies' financial standing and industry positioning, thereby augmenting the dataset's comprehensiveness and analytical utility.

This study selected the question, answer, and evidence_text columns from the dataset for analysis. These columns were considered crucial because of their fundamental roles in enabling the assessment and validation of large language models (LLMs) in tasks related to financial question-answering. The question column provides the textual representation of inquiries within the dataset, serving as the foundation for model input. Concurrently, the answer column contains the corresponding responses generated by the LLMs, enabling model performance and ac-

Table 4.1: Question difficulty categorizations for FinanceBench [52]

FinanceBench	Category Definition	# of Questions
0-RETRIEVE	Retrieve a single data point	53
1-COMPARE	Compare a small number of retrievable data points	18
2-CALC-CHANGE	Calculate relative change in same retrievable data point over time	8
3-CALC-COMPLEX	Calculate complex financial metrics involving multiple data points	47
4-CALC-AND-JUDGE	Calculate complex financial metrics and judge their goodness/healthiness	9
5-EXPLAIN-FACTORS	Explain major driving factors behind a change	2
6-OTHER-ADVANCED	Answer an unusually tricky financial question	4

Table 4.2: Question difficulty categorizations for Test Set

Test Set	Category Definition	# of Questions
0-RETRIEVE	Retrieve a single data point	13
1-COMPARE	Compare a small number of retrievable data points	6
2-CALC-CHANGE	Calculate relative change in same retrievable data point over time	7
3-CALC-COMPLEX	Calculate complex financial metrics involving multiple data points	15
4-CALC-AND-JUDGE	Calculate complex financial metrics and judge their goodness/healthiness	5
5-EXPLAIN-FACTORS	Explain major driving factors behind a change	2
6-OTHER-ADVANCED	Answer an unusually tricky financial question	2

curacy assessment. Additionally, the evidence_text column offers contextual background and supporting information extracted from the source documents, aiding in validating and interpreting model-generated responses. The dataset's source from Hugging Face[53] ensures access to a reputable and widely utilized platform for machine learning resources, thereby enhancing the credibility and reliability of the dataset for research purposes.

4.4 Methodology

The study delves into the efficacy of large language models (LLMs) in addressing financial question-answering tasks, employing a multifaceted methodology encompassing zero-shot learning, few-shot learning, and supervised fine-tuning.

4.4.1 Zero Shot Prompt Engineering

Zero-shot prompt engineering represents a methodology employed to direct large language models (LLMs) in formulating responses to queries without necessitating specific fine-tuning for the given task (Kojima et al., 2022)[54]. This strategy entails furnishing the model with broad instructions or prompts to steer its response generation process. This enables it to deduce answers solely based on the provided context without requiring supplementary task-specific training. In the context of financial question answering, zero-shot prompt engineering plays a crucial role in ensuring that LLMs provide accurate and relevant responses to queries related to financial documents and data. By providing clear instructions to the model regarding the nature of the task and the expected behavior, zero-shot prompts help mitigate the risk of generating inaccurate or irrelevant responses, particularly in complex domains such as finance where precision and accuracy are paramount. The system and user prompt structure used in the LLaMA-2 model follows a specific format to guide the response generation process. The system prompt, enclosed within <s> and </s> tags, provides the model with response instructions. This entails delineating the model's function as a financial chatbot trained to address inquiries derived from the provided information, along with directives for sourcing responses directly from the evidence_text (context) while refraining from incorporating external information or interpretations not explicitly articulated within the provided evidence_text (context). The user prompt, which follows the system prompt, contains the actual query or question posed to the model, along with any additional context or input sections that may aid in generating a response. The complete prompt, encompassing both the system and user prompts, is enveloped within instruction brackets [INST] and [/INST] to signal to the model that the entirety constitutes input. Following the instructions, the model is prompted to generate an output response.[55] The template was populated with each query and its corresponding evidence_text(context) from the dataset during implementation. The resulting prompts were then tokenized and fed into the LLaMA-2 model for response generation. Both the LLaMA-2 and Gemma models adopt a structured system and user prompt format to facilitate the response generation process. The system prompt, delineated within <start_of_turn> and <end_of_turn> tags, provides the model with specific instructions for generating responses Gemma's response generation process follows a similar pattern to that of LLaMA-2, with the model deriving answers from the provided context and adhering to the specified instructions outlined in the system prompt.[56]

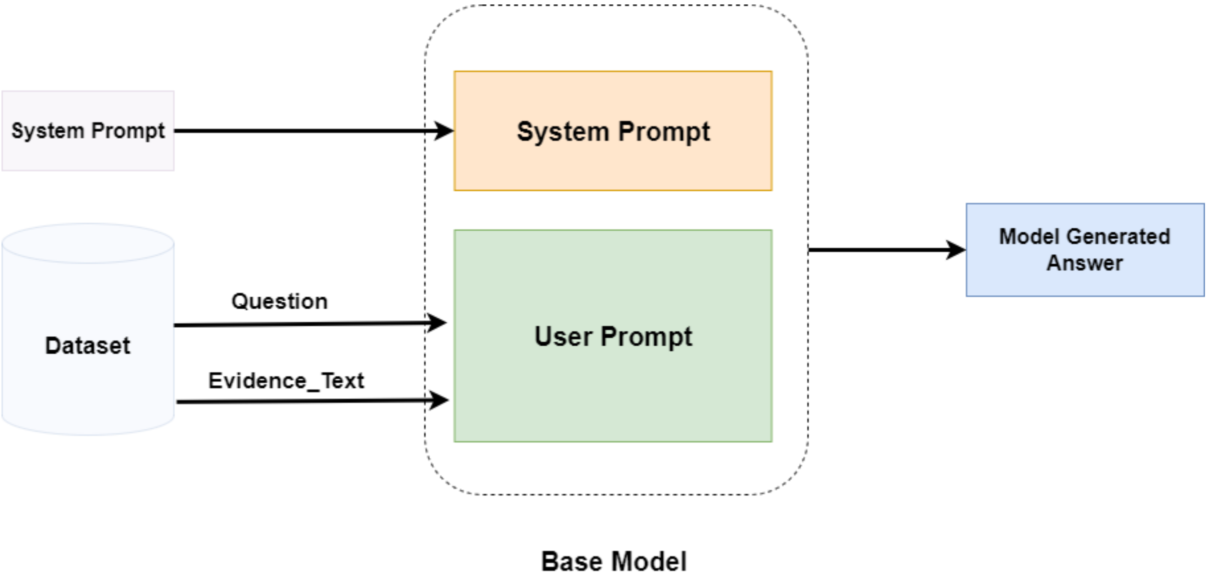


Figure 4.2: Zero Shot Prompt Engineering

4.4.2 Few Shot Prompt Engineering

Few-shot prompt engineering is an advanced method employed in fine-tuning large language models (LLMs) for particular tasks, achieved by furnishing the model with a restricted number of instances or ”shots” about the given task. (Ye et al., 2022)[57] In contrast to zero-shot prompt engineering, which depends solely on general instructions or prompts to direct the model's response generation process, few-shot prompt engineering offers the advantage of integrating task-specific examples to augment the model's comprehension and efficacy. This method empowers LLMs to utilize contextually relevant examples throughout the fine-tuning procedure, enhancing their capacity to produce precise and contextually fitting responses. The need for few-shot prompt engineering arises from the inherent limitations of zero-shot prompt engineering, particularly in complex domains such as finance. While zero-shot prompts provide general guidance to LLMs, they may lack the specificity and granularity required to address nuanced task requirements effectively. By contrast, few-shot prompts offer a more targeted approach by supplying the model with task-specific examples, allowing it to learn from explicit task instances. This not only aids in fine-tuning the model's parameters but also enhances its ability to generalize and adapt to diverse inputs within the specified task domain. During the implementation phase, a few-step prompt engineering techniques were utilized for fine-tuning the LLaMA-2 and Gemma models for financial question-answering tasks. The process involved providing each model with 3 task-specific examples comprising questions, contexts, and corresponding answers. These examples served as training data for the models, allowing them to learn from explicit instances of the task domain. To begin, several carefully curated examples were selected from the dataset, each representing a distinct financial question and its corresponding context. These examples covered various scenarios and query types, ensuring the models were exposed to various aspects of the task domain. Each example consisted of a question posed to the model and a corresponding context extracted from the dataset. Additionally, the correct answer to each question was provided as a reference for the models. These examples were formatted and presented to the models in a structured manner, allowing them to learn from the provided task-specific information effectively.

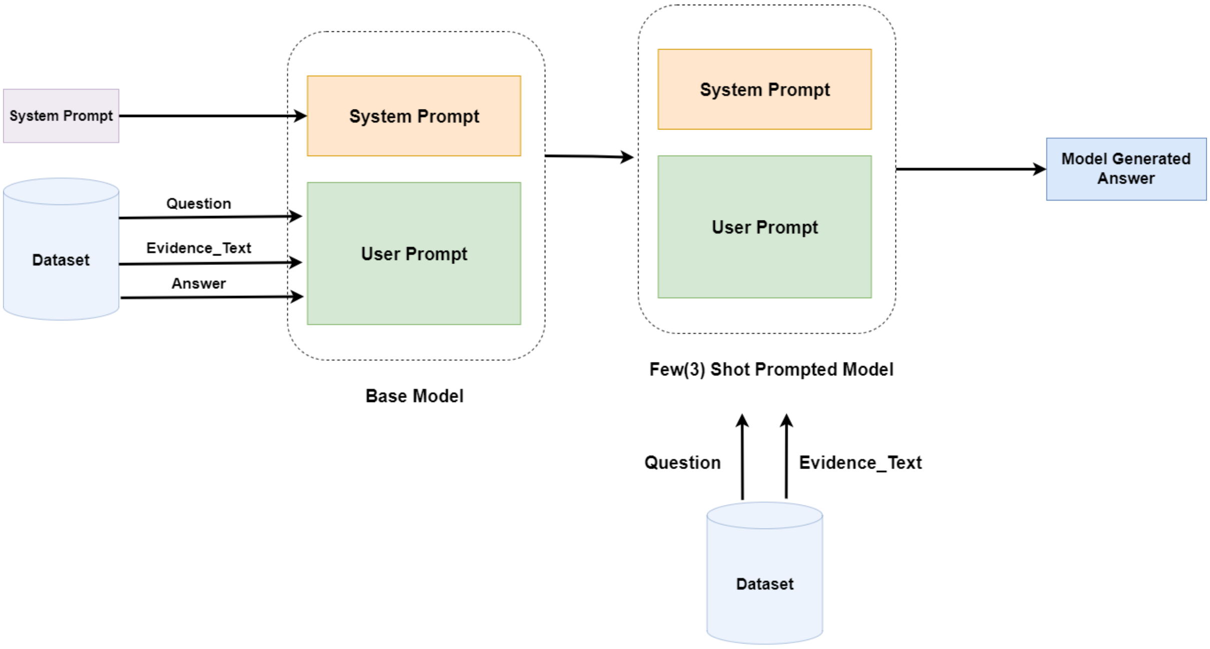


Figure 4.3: Few(3) Shot Prompt Engineering

4.4.3 Supervised Fine Tuning

The approach undertaken in this investigation combines supervised fine-tuning with parameter-efficient fine-tuning methodologies, incorporating the LoRA (Low-Rank Adapters) configuration. The methodology encompasses steps tailored to optimize the efficacy of large language models (LLMs), notably LLaMA-2 and Gemma, for downstream task assignments. The dataset employed in this analysis underwent partitioning into three distinct subsets: an 80-row training set, a 20-row validation set, and a 50-row testing set. This partitioning strategy ensured equitable data distribution for model training, validation, and evaluation purposes. The training set functioned as the principal data reservoir for fine-tuning model parameters, whereas the validation set facilitated iterative model enhancement and hyperparameter optimization to achieve optimal performance. Supervised fine-tuning operations were executed on the training and validation sets to tailor the large language models (LLMs) to the designated task. Supervised fine-tuning encompasses adjusting model parameters using labeled examples sourced from the training data. This process enables the model to acquire task-specific representations and enhance performance metrics through iterative learning. This iterative parameter adjustment and model evaluation process on the validation set ensured that the models achieved optimal performance levels while avoiding overfitting the training data. Fine-tuning was independently performed on the LLaMA-2-7b and Gemma-7b models to evaluate their individual performance in the downstream task. This comparative examination provided insights into the effectiveness of each model and its applicability to the designated domain. The study sought to appraise their capacities to produce precise and contextually appropriate responses to queries within the predefined task parameters through the fine-tuning process applied to both models.

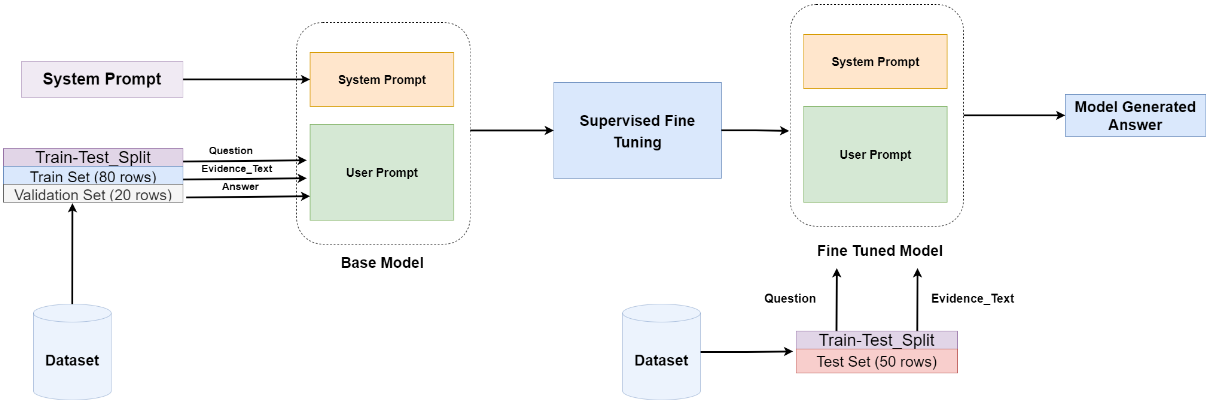


Figure 4.4: Supervised Fine Tuning

4.5 Conclusion

The investigation conducted in this study systematically examined the effectiveness of large language models (LLMs) in tackling financial question-answering (QA) assignments. Employing a multi-phase methodology encompassing zero-shot learning, few-shot learning, and parameter-efficient fine-tuning utilizing LoRA configuration, the study aimed to enhance the models' capabilities for practical financial undertakings. Through extensive experimentation and analysis, several key findings have emerged. Zero-shot prompt engineering provided insights into the models' innate ability to generalize knowledge to unseen tasks, laying the foundation for subsequent fine-tuning phases. Few-shot prompt engineering demonstrated the importance of task-specific examples in enhancing model understanding and performance. Supervised fine-tuning, augmented by parameter-efficient techniques such as LoRA, facilitated iterative model refinement and optimization, improving task performance and generalizability. Additionally, the study highlighted the significance of quantization techniques in enhancing the efficiency and scalability of LLMs for deployment in resource-constrained environments.

Chapter 5

Result & Performance Analysis

5.1 Introduction

The present introductory segment is designed to provide readers with a comprehensive orientation to the fundamental objectives, methodological approaches, and analytical frameworks that underlie the subsequent sections of the document. Particular emphasis is placed on elucidating the significance of these analyses within the specialized finance domain, wherein the precise processing and interpretation of textual data hold substantial implications for decision-making processes and industry applications. Following this contextualization, the introduction proceeds to delineate the overarching structure and thematic focus of the ensuing sections, offering a detailed preview of the key topics to be explored. These include delineating the experimental setup, thoroughly examining train-validation loss dynamics, an overview of evaluation metrics, meticulous model performance comparisons, and in-depth discussions. Through this methodical roadmap, the introduction primes readers to systematically explore and critically evaluate the empirical findings and analytical insights presented throughout the document. It underscores the methodological rigor and scholarly integrity that underpin the subsequent analyses, affirming a steadfast commitment to established academic conventions and best practices.

5.2 Experimental Setup

The hardware configuration delineated herein underscores the foundational importance of robust computing infrastructure in facilitating sophisticated computational tasks. Akin to the backbone of any computational endeavor, the selection and optimization of hardware components are pivotal in ensuring seamless execution and optimal performance of machine learning algorithms and simulations. The amalgamation of powerful processors, ample RAM, and high-performance GPUs is instrumental in tackling complex computations, from training intricate deep-learning models to processing vast datasets. Such hardware configurations enhance efficiency and empower researchers to push the boundaries of computational science and artificial intelligence.

Table 5.1: Hardware Specification

Component	**Specification**
Processor	Intel Xeon CPU
Cores	2 (4 virtual cores)
Speed	2GHz
RAM	31.36 GB
GPU	Tesla P100-PCIE
VRAM	16GB
Disk Space	20GB

The hyperparameters delineated in the table constitute a fundamental aspect of the experimental setup, embodying the nuanced configurations instrumental in the training and refinement of machine learning models. As the extant literature acknowledges, hyperparameters wield considerable influence over model behavior, shaping their capacity to learn intricate patterns and generalize across diverse datasets. A comprehensive understanding and judicious selection of hyperparameters thus underpin the efficacy and reproducibility of experimental outcomes within the realm of natural language processing research.

Table 5.2: Hyperparameters for the models

Hyperparameters	**Value**
lora_config (r)	16
lora alpha	64
lora_dropout	0.1
per_device_train_batch_size	4
gradient_accumulation_steps	4
optimizer	paged_adamw_32bit
learning_rate	2e-5
num_train_epochs	5
evaluation_strategy	steps
eval_steps	0.2
lr_scheduler_type	cosine

5.3 Train-Validation Loss for Supervised Fine Tuning

The comparison between train and validation loss is pivotal in assessing machine learning models' performance and generalization capabilities. Throughout the training process, the primary objective is minimizing the loss function, quantifying the disparity between predicted and actual values. The training loss reflects the model's performance on the training dataset, indicating how well it fits the data. As the model learns from the training data, the training loss typically decreases over epochs, illustrating improved convergence towards optimal parameters. In contrast, the validation loss is computed using a separate dataset not involved in model training. This serves as a proxy for assessing the model's ability to generalize to unseen data. High validation loss may indicate overfitting, where the model has learned to memorize the training data rather than capture underlying patterns. Conversely, consistent or decreasing validation loss suggests that the model generalizes well to new data, reinforcing its robustness.

This section shows training and validation loss for Llama2-7b and Gemma-7b models for supervised fine-tuning.

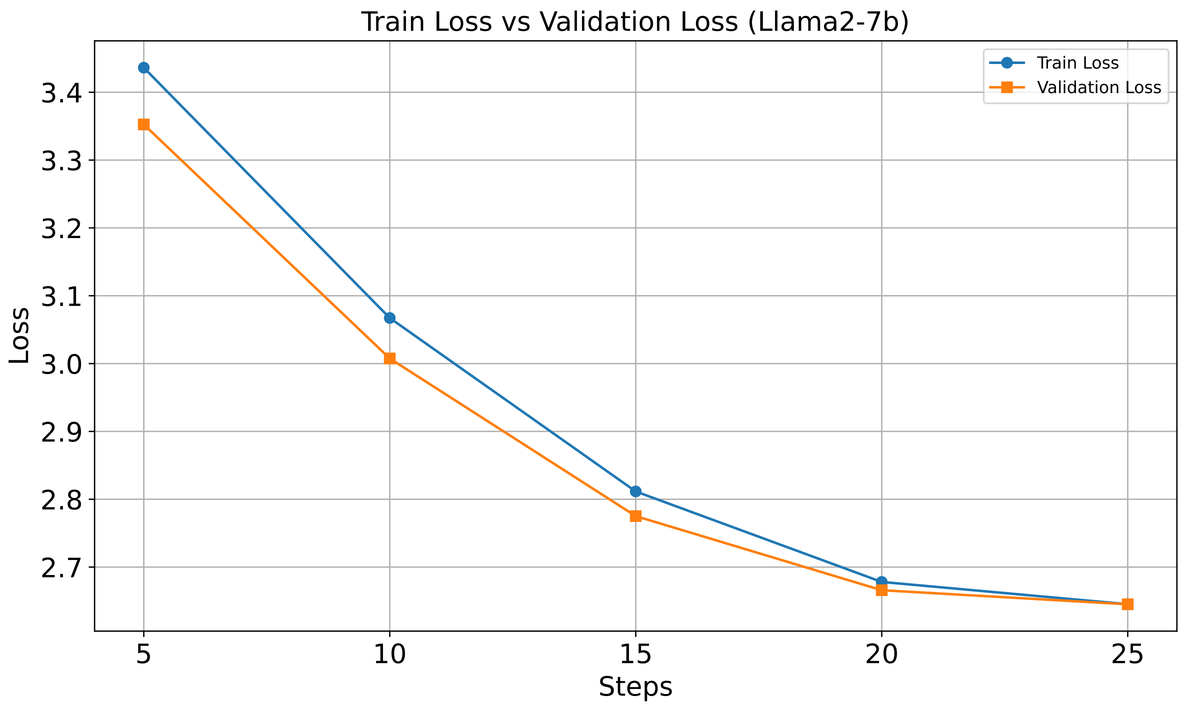


Figure 5.1: Train loss vs Validation loss for Llama2-7b

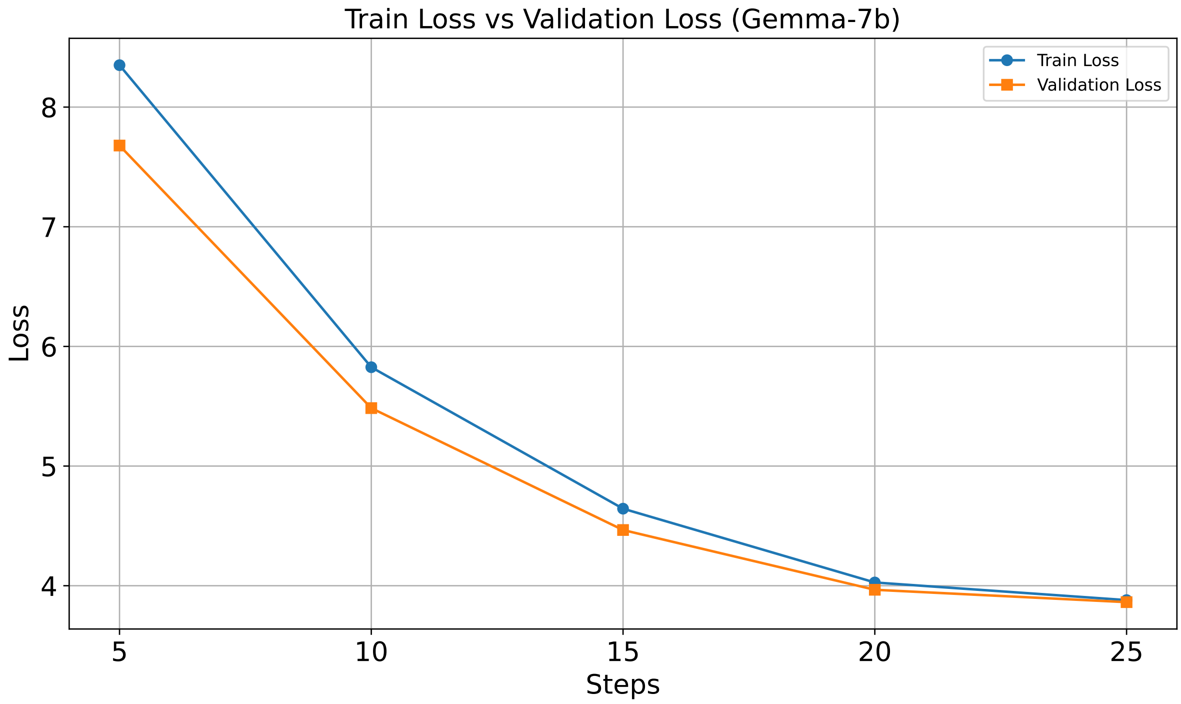


Figure 5.2: Train loss vs. Validation loss for Gemma-7b

5.4 Evaluation Metrics

The evaluation of natural language processing (NLP) models, particularly within finance datasets, stands as a pivotal aspect in assessing their efficacy and reliability. In financial contexts, where precision and accuracy are paramount, the capability of NLP models to furnish correct answers assumes critical significance. Human evaluation serves as a robust benchmark for gauging the performance of these models, providing qualitative insights into their ability to generate responses aligned with factual accuracy and domain-specific knowledge. This comprehensive evaluation process encompasses three primary facets: assessing correctness, identifying instances of hallucination, and analyzing failures in response generation. Through meticulous examination and interpretation of these evaluation metrics, researchers endeavor to refine and optimize NLP models, thereby enhancing their proficiency in addressing the complexities inherent in financial question-answering tasks.

5.4.1 Rouge-L-Score

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence) is a metric commonly employed in natural language processing (NLP) and text summarization tasks to assess the quality of generated text summaries or translations by comparing them against reference texts. It evaluates the overlap between the generated and reference text, emphasizing the longest common subsequences (LCS) between them. ROUGE-L operates under the premise that longer common subsequences indicate better content overlap, thereby reflecting the summary's fidelity to the reference text. ROUGE-L facilitates comparative analyses and model optimisation by quantifying the similarity between the generated output and the reference.[58]

Mathematically, the ROUGE-L score is computed based on the precision, recall, and F1-score metrics, which are calculated as follows:

$$Precision = \frac{\text{LCS}(C, R)}{\text{length}(C)}$$

$$Recall = \frac{\text{LCS}(C, R)}{\text{length}(R)}$$

$$F1\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where C represents the candidate text, R denotes the reference text, and LCS(C, R) signifies the length of the longest common subsequence between the candidate and reference text.[59]
For instance, the candidate text: ”The study examines the impact of climate change on biodiversity,” and the reference text: ”The research investigates how climate change affects biodiversity loss.” The longest common subsequence between these two texts is ”climate change affects biodiversity.” ROUGE-L computes precision, recall, and subsequently, the F1-score by comparing the lengths of the common subsequence to the candidate and reference summaries.
In the context of evaluating question-answering tasks, ROUGE-L can be adapted to assess the similarity between generated answers and ground truth responses. Given a set of questions and corresponding reference answers, ROUGE-L measures the extent to which the generated answers capture the essential information present in the reference responses. This facilitates the quantitative evaluation of question-answering models, allowing researchers to gauge their effectiveness in providing accurate and informative answers.

5.4.2 Cosine Similarity

The cosine similarity metric is a fundamental tool extensively employed in natural language processing (NLP) and information retrieval tasks to quantify the similarity between two vectors in a multi-dimensional space. It assesses the cosine of the angle between the vectors, providing a measure of their directional alignment. This metric operates on the premise that higher cosine similarity scores indicate greater similarity between the vectors, while lower scores imply dissimilarity. Cosine similarity is preferred in NLP tasks due to its ability to measure semantic similarity between textual documents, regardless of their lengths, making it particularly suitable for comparing documents of varying sizes.
Mathematically, the cosine similarity score between two vectors $\vec{A}$ and $\vec{B}$ is computed as the dot product of the vectors divided by the product of their magnitudes:

$$\text{cosine\_similarity}(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{||\vec{A}||||\vec{B}||}$$

where $\cdot$ denotes the dot product and $||\vec{A}||$ represents the magnitude (or Euclidean norm) of vector $\vec{A}$.
For instance, in a hypothetical scenario where $\vec{A} = [1, 2, 3]$ and $\vec{B} = [2, 3, 4]$, the cosine similarity score between these vectors can be calculated using the aforementioned formula. This

computation quantifies the directional alignment between the vectors, offering insight into their semantic similarity.[60]
In the context of evaluating question-answering tasks, cosine similarity is a crucial metric for assessing the resemblance between generated and ground truth responses. Researchers can gauge the accuracy and relevance of the model's outputs by comparing the cosine similarity scores of generated answers with those of reference responses. This quantitative evaluation enables researchers to refine and optimize question-answering systems, ensuring their effectiveness in providing informative and contextually appropriate answers.

5.4.3 Human Evaluation

Evaluation of LLMs, particularly in finance datasets, underscores the paramount importance of human assessment due to the critical nature of obtaining accurate answers. In financial domains, where decisions are often based on precise information, the ability of a model to provide correct answers is of utmost significance. While metrics like ROUGE-L score and cosine similarity are commonly used in natural language processing tasks, they may not be suitable for evaluating the nuanced complexities inherent in financial question answering. These metrics primarily assess the similarity between generated and reference texts, overlooking crucial aspects such as factual correctness and domain relevance. In contrast, human evaluation allows for the qualitative assessment of model performance, providing insights into generated responses' semantic accuracy and contextual appropriateness. Human evaluation serves as a robust benchmark for assessing the effectiveness of language models in generating responses that align with factual accuracy and domain-specific knowledge. The evaluation process typically involves three primary sections: correct answers (correctness), incorrect answers (hallucination), and instances where the model fails to provide any answer (failure). Each section provides valuable insights into the model's performance and its capability to handle various nuances and complexities within the finance domain.

- **Correctness:** The correctness section focuses on assessing the model's ability to generate accurate and relevant answers to given questions. Human evaluators compare the responses generated by the model against ground truth answers to determine their correctness. This section helps gauge the model's proficiency in comprehending the semantics of financial queries and extracting pertinent information from the dataset. By analyzing the correctness of responses, researchers can identify areas of strength and weakness in the model's understanding of financial concepts and its ability to provide precise answers.

- **Hallucination:** The hallucination section addresses instances where the model generates incorrect or irrelevant answers. These errors, often called hallucinations, may stem from misinterpretations of the input query or erroneous information retrieval from the dataset. Evaluators scrutinize the generated responses to identify discrepancies and assess how much the model deviates from the ground truth. Understanding the prevalence and nature of hallucinations is crucial for refining the model's architecture and training procedures to minimize such errors in future iterations.
- **Failure:** In cases where the model fails to provide any answer, the failure section sheds light on the limitations and challenges faced by the model in handling specific queries or scenarios. Failures may arise due to the question's complexity, the dataset's ambiguity set, or gaps in the model's knowledge base. Evaluators analyze these instances to discern patterns and underlying factors contributing to the model's inability to generate responses. Insights from failure analysis inform strategies for enhancing the model's coverage and robustness, improving its overall performance in financial question-answering tasks.

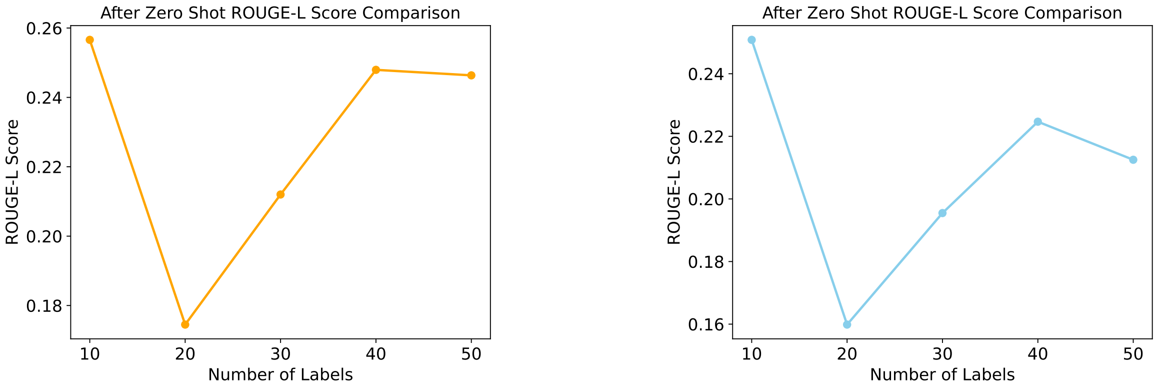
The use of human evaluation metrics allows for a comprehensive comparison of language models' question-answering capabilities in finance datasets. Human evaluation provides qualitative insights into the nature of errors and challenges encountered, guiding efforts to refine and optimize NLP models for specific domains like finance. Researchers can iteratively improve language models' accuracy, relevance, and reliability in addressing complex financial queries and supporting decision-making processes through meticulous analysis and interpretation of evaluation results.

5.5 Model Performance

The following section assesses the performance of the Gemma and LLama2 models in question-answering tasks, employing a comprehensive evaluation framework comprising ROUGE-L, cosine similarity, and human evaluation metrics. The objective is to analyze the models' proficiency in generating accurate and contextually relevant responses, particularly in the domain of finance. The evaluation strategy involves computing ROUGE-L and cosine similarity scores at five distinct evaluation points, each representing an incremental increase in the sample size by 10. Additionally, human evaluators assess the correctness, relevance, and coherence of the generated responses to provide qualitative insights into the models' performance. The evaluation results are presented graphically, facilitating comparative analysis of the model's performance under zero-shot, few-shot, and supervised fine-tuning scenarios. Through this integrated approach, we aim to comprehensively understand the models' question-answering capabilities and identify areas for improvement.

5.5.1 ROUGE-L score

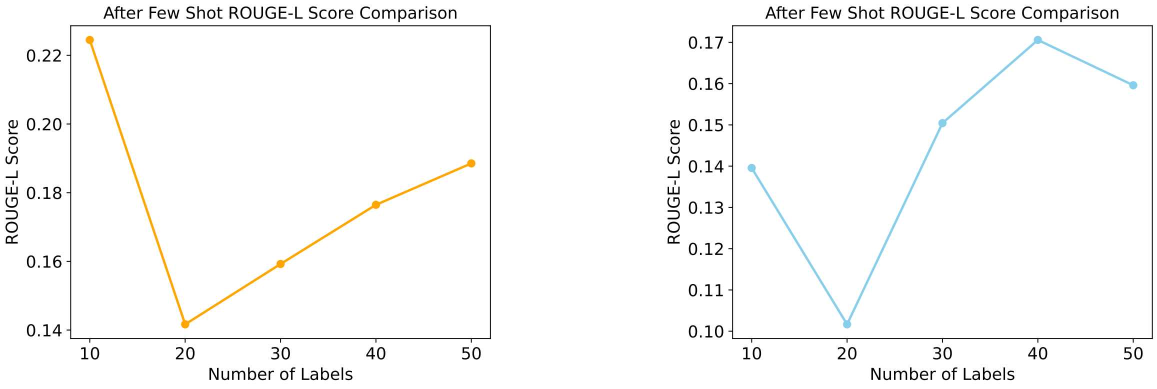
In the section, ROUGE-L score was calculated for Zero Shot Prompt Enginnering, Few(3) Shot Prompt Engineering and Supervised Fine Tuning.



(a) Zero Shot ROUGE-L score for Llama2-7b

(b) Zero Shot ROUGE-L score for Gemma-7b

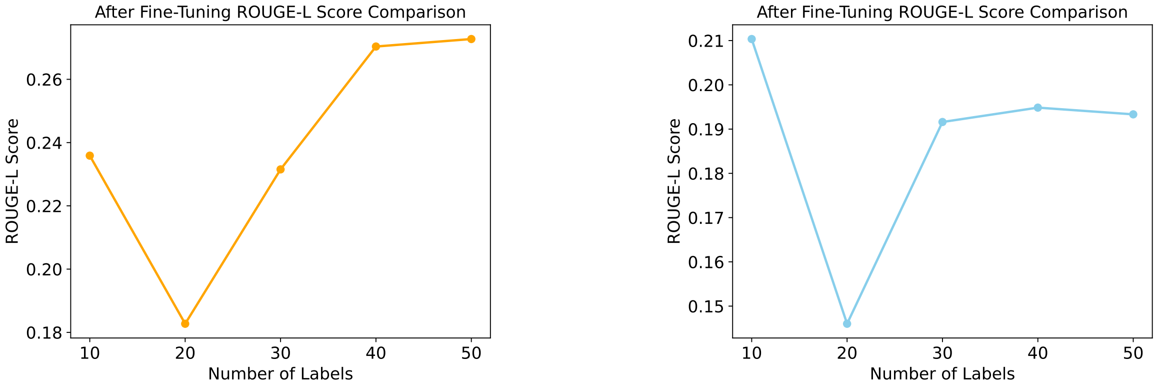
Figure 5.3: Zero Shot Prompt Engineering



(a) Few(3) Shot ROUGE-L score for Llama2-7b

(b) Few(3) Shot ROUGE-L score for Gemma-7b

Figure 5.4: Few(3) Shot Prompt Engineering



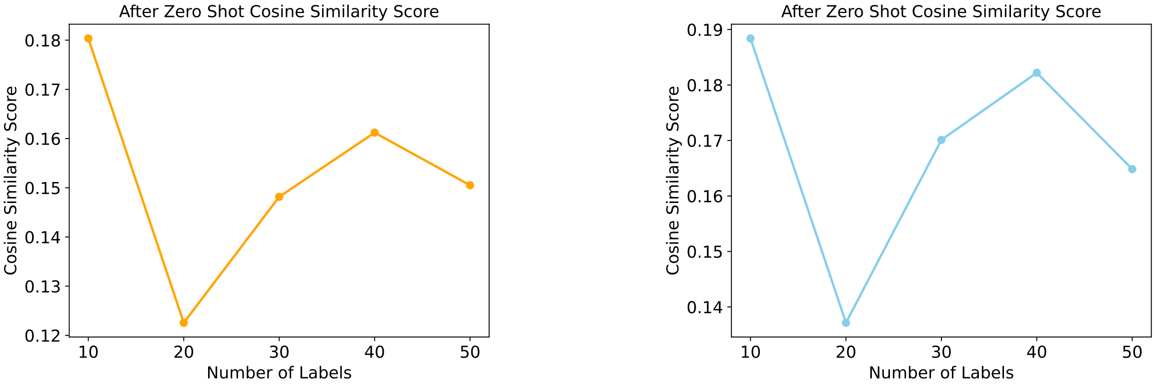
(a) Supervised Fine Tuning ROUGE-L score for Llama2-7b

(b) Supervised Fine Tuning ROUGE-L score for Gemma-7b

Figure 5.5: Supervised Fine Tuning

5.5.2 Cosine similarity score

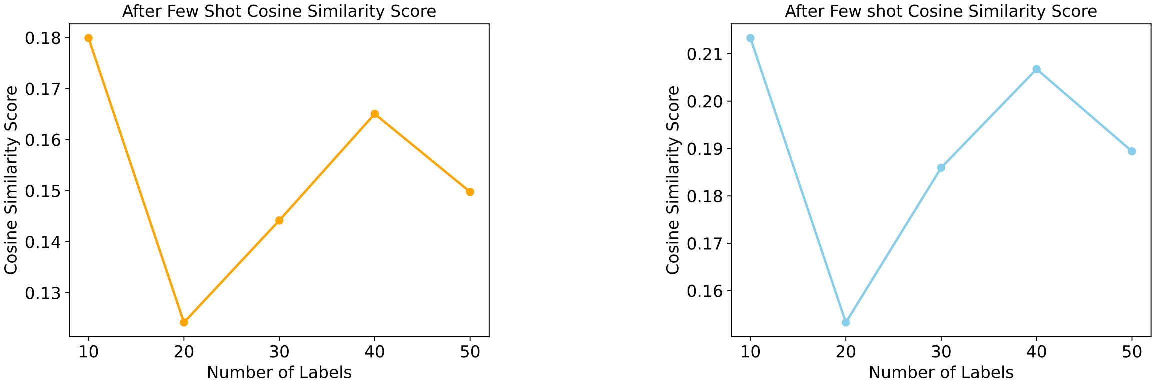
In the section, the Cosine Similarity score was calculated for Zero-Shot Prompt Engineering, Few(3) Shot Prompt Engineering, and Supervised Fine Tuning.



(a) Zero Shot cosine similarity score for Llama2-7b

(b) Zero Shot cosine similarity score for Gemma-7b

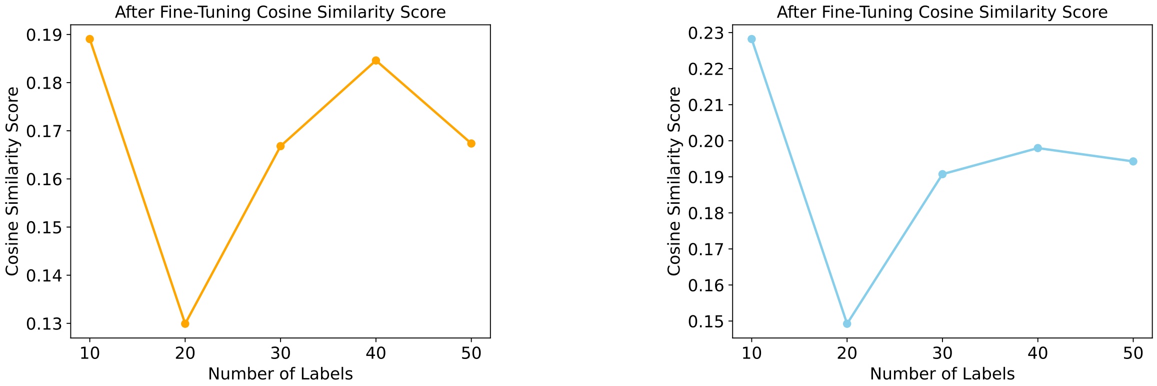
Figure 5.6: Zero Shot Prompt Engineering



(a) Few(3) Shot cosine similarity score for Llama2-7b

(b) Few(3) Shot cosine similarity score for Gemma-7b

Figure 5.7: Few(3) Shot Prompt Engineering



(a) Supervised Fine Tuning cosine similarity score for Llama2-7b

(b) Supervised Fine Tuning cosine similarity score for Gemma-7b

Figure 5.8: Supervised Fine Tuning

5.5.3 Human Evaluation

In this section, a comprehensive human evaluation was conducted to assess the performance of the models in generating answers. Table 5.3 presents the performance evaluation of human assessment across different techniques to illustrate the effectiveness of Llama2 and Gemma in providing correct answers, incorrect answers, and instances where no response was generated denoted as failed to answer. This evaluation offers valuable insights into the models' proficiency across various answer types, contributing to understanding their capabilities in addressing financial queries.

Table 5.4 provides a detailed evaluation of correctness measurement for supervised fine-tuning across different question difficulties. The table categorizes questions into ”Easier” and ”Hard” questions, highlighting the performance of Llama2 and Gemma. Notably, it showcases the models' varying success rates in addressing specific question types, shedding light on their effectiveness in navigating different levels of complexity within financial contexts.

Table 5.3: Performance of Human Evaluation

Zero Shot Prompt Engineering		
Answer-Type	**Llama2**	**Gemma**
Correct Answer	15	5
Incorrect Answer	29	10
Failed to Answer	6	35
Few(3) Shot Prompt Engineering		
Answer-Type	**Llama2**	**Gemma**
Correct Answer	11	3
Incorrect Answer	30	8
Failed to Answer	9	39
Supervised Fine Tuning		
Answer-Type	**Llama2**	**Gemma**
Correct Answer	21	7
Incorrect Answer	29	10
Failed to Answer	0	33

Table 5.4: Correctness Evaluation for different question difficulties for Supervised Fine Tuning

Easier Questions		
Question Difficulty	Llama2	Gemma
0-RETRIEVE	61%	7%
1-COMPARE	66%	0%
2-CALC-CHANGE	14%	0%
Hard Questions		
Question Difficulty	Llama2	Gemma
3-Calc-Complex	33%	20%
4-Calc-And-Judge	20%	20%
5-Explain Factors	50%	100%
6-Other-Advanced	50%	0%

5.6 Comparison of the performances

This section presents comprehensive comparisons between LLama2-7b and Gemma-7b across various evaluation metrics. Both models were scrutinized for their performance regarding ROUGE-L score, cosine similarity, and human evaluation. The analysis delves into their strengths and weaknesses, highlighting their capabilities and limitations in question-answering tasks.

5.6.1 ROUGE-L score

For the zero-shot scenario, LLama2 consistently achieves higher ROUGE-L scores than Gemma across multiple epochs. This suggests that LLama2 exhibits a stronger inherent capability to generate summaries without prior training on specific prompts. Conversely, Gemma demonstrates comparatively lower ROUGE-L scores in the zero-shot setting, indicating a lesser proficiency in generating accurate summaries without task-specific training. In the few-shot scenario, LLama2 outperforms Gemma, albeit with a narrower margin than the zero-shot setting. Despite both models benefiting from additional prompt information, LLama2 maintains its superiority in ROUGE-L scores, indicating its greater adaptability and effectiveness in leveraging limited prompt information for question-answering. In the supervised fine-tuning phase, both models show improvements in ROUGE-L scores compared to their zero-shot and few-shot counterparts. However, LLama2 consistently maintains higher ROUGE-L scores than Gemma across epochs, indicating its superior performance even after fine-tuning task-specific data. This suggests that LLama2 can leverage task-specific information for answer generation, leading to more accurate and relevant summaries than Gemma.

5.6.2 Cosine similarity score

For the zero-shot scenario, both models exhibit varying degrees of cosine similarity. While LLama2 achieves higher cosine similarity scores, indicating closer alignment between the generated responses and reference answers, Gemma demonstrates relatively lower scores. This suggests that LLama2 outperforms Gemma in capturing semantic similarity without the need for additional training data. In the few-shot setting, LLama2 continues to maintain a lead in cosine similarity scores compared to Gemma. However, the gap between the two models diminishes, indicating that Gemma's performance improves by incorporating limited training examples. Despite this improvement, LLama2 remains superior in preserving semantic similarity across responses. In the supervised fine-tuning phase, LLama2 consistently demonstrates superior cosine similarity scores compared to Gemma. This suggests that LLama2 benefits more from fine-tuning on annotated training data, further enhancing its ability to generate responses that closely resemble the reference answers.

5.6.3 Human Evaluation

The Llama2 model demonstrates a higher accuracy in generating correct answers than Gemma. When evaluated under the zero-shot prompt engineering technique, Llama2 consistently outperforms Gemma regarding correct answer generation. Specifically, Llama2 achieves higher percentages (30%) of correct answers than Gemma (18%), indicating its proficiency in comprehending queries and extracting relevant information from the dataset. However, it is important to note that Llama2 also produces more incorrect answers (58%) than Gemma (22%). Despite this, the overall accuracy of Llama2 in providing correct responses surpasses that of Gemma. For failure of the answer, Llama2 outperforms Gemma by the failure of (12%) whereas for Gemma it was (60%) Under the few-shot prompt engineering technique, Llama2 and Gemma exhibit contrasting performance patterns. While Llama2 maintains a relatively high percentage of correct answers(22%) than Gemma (6%). But it generated many incorrect responses (60%). In contrast, Gemma produces fewer incorrect answers (16%) but struggles with responding to a larger number of queries (78%), whereas for llama2, it was lower (18%), resulting in a higher rate of unanswered questions. This highlights the trade-off between accuracy and completeness in question-answering tasks, with Llama2 prioritizing accuracy and Gemma emphasizing completeness.

In the supervised fine-tuning technique, Llama2-7b and Gemma-7b models undergo refinement to improve their performance. Llama2 demonstrates superior accuracy in generating correct answers (32%) compared to Gemma (20%), reflecting its enhanced comprehension of queries and extraction of relevant information. However, like previous techniques, Llama2 exhibits a higher frequency of incorrect answers(64%). Conversely, Gemma minimizes the occurrence of incorrect responses (16%) but struggles with providing responses to a significant number of queries (64%), indicating limitations in its comprehension abilities. For Llama2, it was lower (4%).

Table 5.4 provides a comprehensive overview of the correctness evaluation for different question difficulties in supervised fine-tuning. Llama2 demonstrates relatively higher accuracy rates for easier questions across all question types than Gemma. Specifically, for question difficulties involving information retrieval, Llama2 exhibits notably higher performance, with accuracy rates of 61%. Conversely, Gemma shows lower accuracy rates for easier questions, particularly for information retrieval, where it achieved 7% accuracy. Llama2 and Gemma display decreased accuracy rates compared to easier questions in the case of harder questions. For questions such as calculating complex financial metrics, Llama2 achieves a 33% accuracy rate, relatively higher than Gemma, which achieved 20%.

Considering these observations, the evaluation of Llama2 and Gemma highlights the trade-offs between accuracy and completeness in question-answering tasks. Llama2 is more accurate in generating correct answers but is prone to providing more incorrect responses. On the other hand, Gemma minimizes the occurrence of incorrect answers but struggles with providing responses to a substantial number of queries. In comparing the performance of the zero-shot, few-shot, and Supervised Fine Tuning (SFT) techniques for both the Llama2 and Gemma models, a notable trend emerges regarding the distribution of responses. Across both models, the few-shot technique consistently yields fewer responses than the zero-shot technique. This observation holds true for all evaluation metrics, including correctness, incorrectness, and failures to provide answers. The discrepancy in the number of responses between the zero-shot and few-shot techniques suggests that the few-shot approach may be inherently more selective or conservative in generating responses. This could be attributed to the limited amount of additional training data provided during the few-shot learning process, resulting in a more cautious approach to answer generation.

5.7 Discussion

In evaluating various metrics such as ROUGE-L score, cosine similarity, and human evaluation, the focus primarily lies in assessing the correctness, incorrectness, and failures of the models in generating answers. Specifically, the financial industry demands a nuanced understanding of the trade-offs between sacrificing incorrectness and failing to provide answers. The evaluation highlights the contrasting capabilities of the Llama2-7b and Gemma-7b models in the financial domain. Llama2 demonstrates a higher proficiency in generating correct answers across all evaluated techniques, including zero-shot, few-shot, and Supervised Fine Tuning (SFT), than Gemma. However, LLama2's tendency to produce more incorrect answers offset this advantage. On the other hand, Gemma minimizes incorrect answers but faces challenges in responding to a substantial number of queries, resulting in a higher rate of unanswered questions. While LLama2 and Gemma exhibit strengths and weaknesses in generating responses, the finance industry seeks models that balance accuracy and completeness. Current models demonstrate varying degrees of proficiency in understanding and processing financial queries, yet a pressing need remains to enhance their efficiency and accuracy in the financial domain. Moving forward, the consideration shifts towards enhancing the efficiency and effectiveness of question-answering models in the financial sector. To address this, future research should prioritize developing models with improved accuracy in generating responses while minimizing incorrectness and failures to provide answers. Additionally, there is a growing emphasis on enhancing the models' understanding of financial concepts and terminology, enabling them to generate more contextually relevant and accurate responses. Making question-answering models more efficient in the financial domain cannot be overstated. As financial institutions increasingly rely on AI-powered systems for tasks such as customer support, investment analysis, and risk management, the accuracy and reliability of these models become paramount. Inaccurate or incomplete responses can have significant implications for financial decision-making, potentially leading to financial losses or reputational damage for institutions. Therefore, there is a critical need to continuously refine and improve question-answering models to meet the evolving demands of the financial industry and ensure their effectiveness in real-world applications.

5.8 Conclusion

The evaluation of LLama2-7b and Gemma-7b models across various metrics, including ROUGE-L score, cosine similarity, and human evaluation, provides valuable insights into their respective strengths and weaknesses in addressing question-answering tasks, particularly within the financial domain. While LLama2-7b demonstrates a higher proficiency in generating correct answers across different techniques, including zero-shot, few-shot, and Supervised Fine Tuning (SFT), Gemma-7b exhibits a contrasting pattern by minimizing incorrect answers but faces challenges in providing responses to a significant number of queries. Future research endeavors should prioritize enhancing question-answering models' efficiency and effectiveness in the financial sector, improving accuracy while minimizing incorrectness and failures to provide answers.

Chapter 6

Conclusion & Future Works

6.1 Introduction

This chapter offers a succinct summary of the entire study, encompassing the research scope, related literature, novel insights, experimental findings, and concluding remarks. Furthermore, it offers a brief glimpse into potential avenues for future research endeavors.

6.2 Summary

In conclusion, the evaluation of Llama2-7b and Gemma-7b models in the context of financial question-answering tasks sheds light on their respective capabilities and limitations. Key insights were gleaned regarding the models ' performance dynamics through a comprehensive assessment encompassing zero-shot learning, few-shot learning, and supervised fine-tuning methodologies. Llama2-7b exhibited a higher proficiency in generating correct answers across different techniques, including zero-shot, few-shot, and supervised fine-tuning. Despite its accuracy, Llama2-7b tended to produce more incorrect answers, highlighting a trade-off between accuracy and completeness in question-answering tasks. Conversely, Gemma-7b demonstrated a contrasting pattern by minimizing incorrect answers but facing challenges in responding to a substantial number of queries. This indicates a trade-off between accuracy and comprehensiveness in response generation. Furthermore, the evaluation revealed nuanced differences in the models' performance metrics, including ROUGE-L score, cosine similarity, and human evaluation. Llama2-7b consistently outperformed Gemma-7b across various evaluation scenarios, underscoring its superior adaptability and effectiveness in leveraging limited prompt information for question-answering. The study highlights the intricate trade-offs inherent in developing question-answering models for the financial domain. While Llama2-7b and Gemma-7b exhibit strengths and weaknesses, the findings underscore the need for continued research efforts to enhance the efficiency and accuracy of large language models in financial question-answering tasks. By addressing these challenges, the field can advance towards more precise and reliable question-answering systems tailored for real-world applications in the financial industry.

6.3 Conclusion

This study systematically evaluated the performance of Llama2-7b and Gemma-7b models in addressing financial question-answering tasks, employing a multi-phase methodology encompassing zero-shot learning, few-shot learning, and supervised fine-tuning. Through meticulous experimentation and analysis, several key findings emerged regarding the models' efficacy and suitability for practical financial applications. The findings underscore the intricate trade-offs in developing question-answering models tailored to the financial domain. While Llama2-7b and Gemma-7b exhibit strengths and weaknesses, the study emphasizes the importance of continued research efforts to enhance the efficiency and accuracy of large language models in financial question-answering tasks. Moving forward, future research endeavors should prioritize refining question-answering models to strike a balance between accuracy and completeness, thereby meeting the evolving demands of the financial industry. By addressing these challenges, the field can advance toward more precise and reliable question-answering systems capable of supporting critical financial decision-making processes.

6.4 Limitation

Despite the insights gained from this research, several limitations warrant acknowledgment. Firstly, the specific models and datasets constrain the study's findings. While Llama2-7b and Gemma 7b represent state-of-the-art LLMs, their performance may not indicate all LLMs in the financial question-answering domain. Moreover, the analysis is based on a single dataset, potentially limiting the generalizability of the results. Future research could benefit from exploring a wider range of LLMs and datasets to better understand their capabilities. Additionally, the study's focus on a specific task within the financial domain may limit its applicability to other domains. Financial question answering represents a subset of NLP tasks, and LLM performance may vary across different domains and tasks. Therefore, caution should be exercised when extrapolating the findings of this research to other contexts. Furthermore, the research is subject to the constraints of the methodologies employed. While zero-shot learning, few-shot learning, and fine-tuning methodologies offer valuable insights into LLM performance, the study's limited focus on hyperparameter tuning and the use of a small dataset of only 100 samples for fine-tuning may impact the generalizability of the findings. Additionally, a pre-trained model loaded on a quantized configuration could limit performance. Overall, while this research provides valuable insights into the efficacy of LLMs in financial question answering, its findings are subject to certain limitations inherent in the methodologies, models, and datasets utilized. Addressing these limitations and conducting further research across diverse LLMs, datasets, and evaluation methods will contribute to a more comprehensive understanding of LLM capabilities in the financial domain.

6.5 Future Works

Future research endeavors in the realm of large language models (LLMs) for financial question answering should strategically address key challenges and explore innovative methodologies to advance the field. Several avenues for future work emerge from the findings and observations of the current study. One area of focus should be investigating the discrepancy between the performance of zero-shot and few-shot learning techniques. Despite the potential benefits of few-shot learning in providing additional task-specific examples, the observed lower performance than zero-shot learning warrants further investigation. Future studies can explore the underlying reasons for this phenomenon and develop strategies to mitigate the performance gap. This may involve analyzing the effectiveness of different prompt engineering techniques and exploring novel approaches to leverage limited training data more effectively. Addressing the challenges associated with model hallucinations and erroneous responses is imperative to enhance the reliability and trustworthiness of LLMs in financial question-answering tasks. Future research should explore various techniques to minimize these issues, including supervised fine-tuning with larger and more diverse datasets. Additionally, unsupervised learning approaches and reinforcement learning with human feedback (RLHF) techniques can be employed to iteratively refine model outputs and reduce the occurrence of hallucinations and inaccuracies.

Moreover, recent advancements such as the retrieval-augmented generation (RAG) technique show promise in improving the quality of model-generated responses by incorporating external knowledge sources. Investigating the applicability of such techniques in the context of financial document translation and summary generation could further enhance the capabilities of LLMs for financial applications. Furthermore, future studies should prioritize exploring the scalability and efficiency of LLMs for real-world deployment in financial settings. This includes optimizing model architectures and training procedures to minimize computational resources while maintaining high-performance levels. Additionally, evaluating the robustness of LLMs to handle large volumes of financial data and adapt to dynamic market conditions is crucial for ensuring their practical utility in financial decision-making processes. Interdisciplinary collaboration between researchers in natural language processing (NLP), finance, and AI ethics is essential to address the ethical and societal implications of deploying LLMs in financial settings. This includes examining issues related to bias, fairness, and transparency in model outputs and ensuring compliance with regulatory standards and guidelines. Moreover, exploring techniques to enhance the interpretability of LLMs and facilitate human-AI collaboration in financial analysis tasks can contribute to building trust and confidence in these AI-powered systems. By pursuing these future research directions, the field can advance toward developing more robust, reliable, and ethically sound LLMs for financial question-answering, ultimately enabling more informed and efficient decision-making processes in the financial sector.

REFERENCES

[1] J. W. Goodell, S. Kumar, W. M. Lim, and D. Pattnaik, “Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis,” *Journal of Behavioral and Experimental Finance*, vol. 32, p. 100577, 2021.

[2] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, “A survey on evaluation of large language models,” vol. 15, no. 3, 2024.

[3] H. Zhao, Z. Liu, Z. Wu, Y. Li, T. Yang, P. Shu, S. Xu, and et al., “Revolutionizing finance with llms: An overview of applications and insights,” *arXiv preprint arXiv:2401.11641*, 2024.

[4] B. Merkus, “An assessment of zero-shot open book question answering using large language models,” Master’s thesis, 2023.

[5] Y. Li, S. Wang, H. Ding, and H. Chen, “Large language models in finance: A survey,” in *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 374–382, 2023.

[6] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, and et al., “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.

[7] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, and et al., “Gemma: Open models based on gemini research and technology,” *arXiv preprint arXiv:2403.08295*, 2024.

[8] Datacamp, “What is Natural Language Processing (NLP)? A Comprehensive Guide for Beginners.” https://www.datacamp.com/blog/what-is-natural-language-processing. Accessed: 12/04/2024.

[9] Deeplearning.ai, "Natural Language Processing." https://www.deeplearning.ai/resources/natural-language-processing/. Accessed: 12/04/2024.

[10] Towards Data Science, "Your Guide to Natural Language Processing (NLP)." https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1. Accessed: 12/04/2024.

[11] Turing, "A Guide on Word Embeddings in NLP." https://www.turing.com/kb/guide-on-word-embeddings-in-nlp. Accessed: 12/04/2024.

[12] Neuroflash, "Large Language Models: Understanding the Future of NLP." https://neuroflash.com/blog/large-language-models-understanding-the-future-of-nlp/. Accessed: 12/04/2024.

[13] Data Science Dojo, "Best Large Language Models (LLMs) in 2024." https://datasciencedojo.com/blog/best-large-language-models/. Accessed: 12/04/2024.

[14] K. LLC, "Gemma vs. llama vs. mistral: A comparative analysis with a coding twist," *Medium*, 2024. Accessed: 12/04/2024.

[15] T. A. Dream, "Google gemma open source llm: Everything you need to know," *The AI Dream*, 2024. Accessed: 12/04/2024.

[16] Hugging Face Inc., "Transformers Documentation: Llama2." `https://huggingface.co/docs/transformers/main/model_doc/llama2`. Accessed: 12/04/2024.

[17] "Making llms even more accessible with bitsandbytes, 4-bit quantization and qlora."

[18] Towards Data Science, "Democratizing llms: 4-bit quantization for optimal llm inference." https://towardsdatascience.com/democratizing-llms-4-bit-quantization-for-optimal-llm-inference-be30cf4e0e34, 2024. Accessed: 12/04/2024.

[19] T. Dettmers, "Bitsandbytes." `https://github.com/TimDettmers/bitsandbytes`, 2024. Accessed: 12/04/2024.

[20] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen, *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 220–235, 2023.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[22] Hatchworks, "LARGE LANGUAGE MODELS: CAPABILITIES, ADVANCEMENTS, AND LIMITATIONS [2024]." https://hatchworks.com/blog/gen-ai/large-language-models-guide/. Accessed: 12/04/2024.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[25] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp. 242–264, IGI global, 2010.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[27] J. Bandy and N. Vincent, "Addressing" documentation debt" in machine learning research: A retrospective datasheet for bookcorpus," *arXiv preprint arXiv:2105.05241*, 2021.

[28] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[29] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[30] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," *arXiv preprint arXiv:1705.03551*, 2017.

[31] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," *arXiv preprint arXiv:1806.03822*, 2018.

[32] B. Kostić, J. Risch, and T. Möller, "Multi-modal retrieval of tables and texts using tri-encoder models," *arXiv preprint arXiv:2108.04049*, 2021.

[33] Q. Zhang, S. Chen, D. Xu, Q. Cao, X. Chen, T. Cohn, and M. Fang, "A survey for efficient open domain question answering," *arXiv preprint arXiv:2211.07886*, 2022.

[34] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "Bloomberggpt: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.

[35] H. Yang, X.-Y. Liu, and C. D. Wang, "Fingpt: Open-source financial large language models," *arXiv preprint arXiv:2306.06031*, 2023.

[36] S. Choi, W. Gazeley, S. H. Wong, and T. Li, "Conversational financial information retrieval model (confirm)," *arXiv preprint arXiv:2310.13001*, 2023.

[37] R. S. Shah, K. Chawla, D. Eidnani, A. Shah, W. Du, S. Chava, N. Raman, C. Smiley, J. Chen, and D. Yang, "When flue meets flang: Benchmarks and large pre-trained language model for financial domain," *arXiv preprint arXiv:2211.00083*, 2022.

[38] F. Zhu, W. Lei, Y. Huang, C. Wang, S. Zhang, J. Lv, F. Feng, and T.-S. Chua, "Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance," *arXiv preprint arXiv:2105.07624*, 2021.

[39] M. Maia, S. Handschuh, A. Freitas, B. Davis, R. McDermott, M. Zarrouk, and A. Balahur, "Www'18 open challenge: financial opinion mining and question answering," in *Companion proceedings of the the web conference 2018*, pp. 1941–1942, 2018.

[40] Z. Chen, W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. Routledge, *et al.*, "Finqa: A dataset of numerical reasoning over financial data," *arXiv preprint arXiv:2109.00122*, 2021.

[41] Z. Chen, S. Li, C. Smiley, Z. Ma, S. Shah, and W. Y. Wang, "Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering," *arXiv preprint arXiv:2210.03849*, 2022.

[42] J. C. S. Alvarado, K. Verspoor, and T. Baldwin, "Domain adaption of named entity recognition to support credit risk assessment," in *Proceedings of the Australasian Language Technology Association Workshop 2015*, pp. 84–90, 2015.

[43] E. Callanan, A. Mbakwe, A. Papadimitriou, Y. Pei, M. Sibue, X. Zhu, Z. Ma, X. Liu, and S. Shah, "Can gpt models be financial analysts? an evaluation of chatgpt and gpt-4 on mock cfa exams," *arXiv preprint arXiv:2310.08678*, 2023.

[44] E. Kamalloo, N. Dziri, C. L. Clarke, and D. Rafiei, "Evaluating open-domain question-answering in the era of large language models," *arXiv preprint arXiv:2305.06984*, 2023.

[45] Z. Zhang, C. Zheng, D. Tang, K. Sun, Y. Ma, Y. Bu, X. Zhou, and L. Zhao, "Balancing specialized and general skills in llms: The impact of modern tuning and data strategy," *arXiv preprint arXiv:2310.04945*, 2023.

[46] H. Zhang, Q. Si, P. Fu, Z. Lin, and W. Wang, "Are large language models table-based fact-checkers?," *arXiv preprint arXiv:2402.02549*, 2024.

[47] H. A. Alawwad, A. Alhothali, U. Naseem, A. Alkhathlan, and A. Jamal, "Enhancing textbook question answering task with large language models and retrieval augmented generation," *arXiv preprint arXiv:2402.05128*, 2024.

[48] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[49] T. Adewumi, N. Habib, L. Alkhaled, and E. Barney, "On the limitations of large language models (llms): False attribution," *arXiv preprint arXiv:2404.04631*, 2024.

[50] P. Islam, A. Kannappan, D. Kiela, R. Qian, N. Scherrer, and B. Vidgen, "Financebench: A new benchmark for financial question answering," *arXiv preprint arXiv:2311.11944*, 2023.

[51] L. Zhang, K. Jijo, S. Setty, E. Chung, F. Javid, N. Vidra, and T. Clifford, "Enhancing large language model performance to answer questions and extract information more accurately," *arXiv preprint arXiv:2402.01722*, 2024.

[52] Z. Nguyen, A. Annunziata, V. Luong, S. Dinh, Q. Le, A. H. Ha, C. Le, H. A. Phan, S. Raghavan, and C. Nguyen, "Enhancing q&a with domain-specific fine-tuning and iterative reasoning: A comparative study," *arXiv preprint arXiv:2404.11792*, 2024.

[53] Patronus AI, "Financebench." https://huggingface.co/datasets/PatronusAI/financebench, 2022. Accessed: 12/04/2024.

[54] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22199–22213, 2022.

[55] Replicate, "How to prompt: Llama." https://replicate.com/blog/how-to-prompt-llama, Year. Accessed: 12/04/2024.

[56] C. Coder, "Getting started with google's gemma llm using huggingface libraries," *Medium*. Accessed: 12/04/2024.

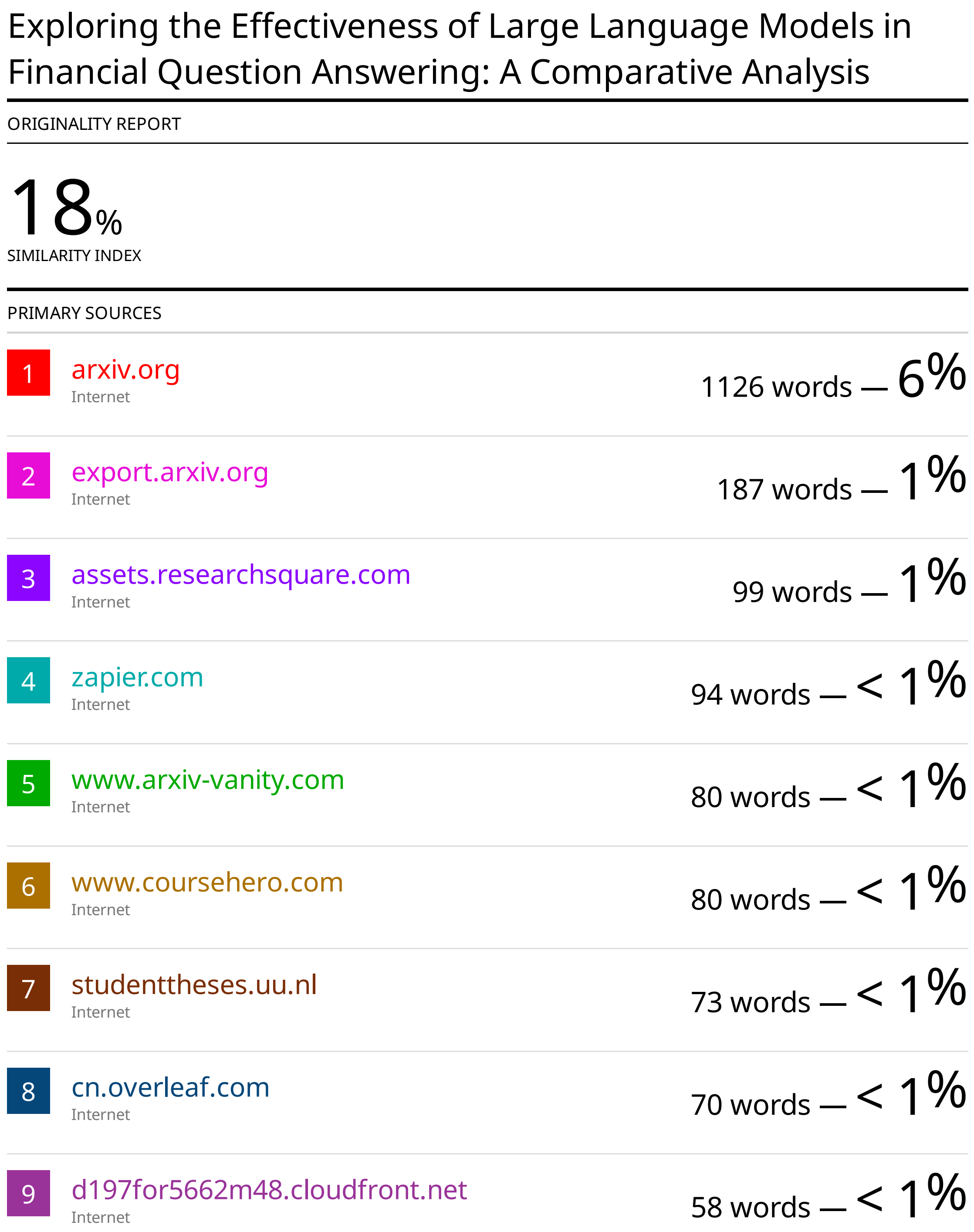
[57] X. Ye and G. Durrett, "The unreliability of explanations in few-shot prompting for textual reasoning," *Advances in neural information processing systems*, vol. 35, pp. 30378–30392, 2022.

[58] FreeCodeCamp, "What is rouge and how it works for evaluation of summaries," *freeCodeCamp.org*, 2019. Accessed: 12/04/2024.

[59] NLPlanet, "Two minutes nlp: Learn the rouge metric by examples," *Medium*, 2021. Accessed: 12/04/2024.

[60] LearnDataSci, "Cosine similarity," *LearnDataSci.com*. Accessed: 12/04/2024.

SIMILARITY INDEX



Exploring the Effectiveness of Large Language Models in Financial Question Answering: A Comparative Analysis

ORIGINALITY REPORT

18%

SIMILARITY INDEX

PRIMARY SOURCES

1	arxiv.org Internet	1126 words — 6%
2	export.arxiv.org Internet	187 words — 1%
3	assets.researchsquare.com Internet	99 words — 1%
4	zapier.com Internet	94 words — < 1%
5	www.arxiv-vanity.com Internet	80 words — < 1%
6	www.coursehero.com Internet	80 words — < 1%
7	studenttheses.uu.nl Internet	73 words — < 1%
8	cn.overleaf.com Internet	70 words — < 1%
9	d197for5662m48.cloudfront.net Internet	58 words — < 1%