ICCA

3rd International Conference on Computing Advancements - ICCA 2024
October 17-18, 2024

Exploring the Effectiveness of Large Language Models in Financial Question Answering

Authors:

Mondol Mridul Provakar
Rajshahi University of
Engineering & Technology
Rajshahi, Bangladesh
mondolmridul007@gmail.com

Emrana Kabir Hashi
Rajshahi University of
Engineering & Technology
Rajshahi, Bangladesh
emranakabir@gmail.com

Department of Computer Science
Faculty of Science and Technology

American
International
University-Bangladesh

Presentation Outline

- Introduction
- Motivation & Objectives
- Literature Review
- Workflow of the Research
- Dataset Description
- Methodology
- Results
- Conclusion
- Future Works
- References

Introduction

Large Language Models:

Definition: A type of artificial intelligence (AI) that are trained on massive amounts of text data.

Characteristics:

- **Large size:** Trained on billions or even trillions of words of text data.
- **Generative capabilities:** Can generate new text, including translations, summaries, question-answering etc.
- **Domain adaptation:** Can be fine-tuned to perform well on specific tasks or domains. Example: medical chatbot, legal document summarizer etc.

Introduction (Cont’d)

Challenges for Large Language Models:

- **Cost:** Training and inferencing can be computationally expensive.
- **Generalization:** Can have difficulty generalizing to new situations or domains.
- **Fairness:** Can be used to create unfair or discriminatory systems.
- **Privacy:** Can be used to infer sensitive information about people.

Introduction (Cont’d)

Table 1: Best Large Language Models in 2024[1]

LLM	Developer	Popular Apps	# of Parameters (billions)	Access
GPT-3	OpenAI	Microsoft, Duolingo, Stripe, Zapier, Dropbox, ChatGPT	175	API
Gemini	Google	Bard & Nano (some queries)	1.8 - 3.25 (varies)	API
PaLM 2	Google	Bard, Docs, Gmail, and other Google apps	340	API
Llama 2	Meta	Undisclosed	7, 13, and 70	Open Source
Claude 2	Anthropic	Slack, Notion, Zoom	Unknown	API
Falcon	Technology Innovation Institute	Undisclosed	1.3, 7.5, 40, and 180	Open Source
MPT	Mosaic	Undisclosed	7 and 30	Open Source
Mixtral	Mistral AI	Undisclosed	46.7	Open Source

Introduction (Cont'd)

Table 2: Model Configuration for Llama2 and Gemma[2]

Configuration	Llama2	Gemma
Vocabulary Size	32,000	256,000
Context Length tokens	4096	8192
Hidden Size	4,096	3,072
Number of Hidden Layers	32	28
Number of Attention Heads	32	16

Motivations & Objectives

Motivation:

- AI innovations are transforming the financial industry.
- Large Language Models (LLMs) offer promising opportunities in financial analysis.
- Current LLMs show potential but face challenges in addressing financial questions.

Objective:

- Assess proficiency of LLMs, specifically Llama2-7b and Gemma-7b, in financial question answering.
- Evaluate model performance using metrics like ROUGE-L, cosine similarity and human evaluation.
- Identify strengths and limitations of LLMs in financial tasks.

Literature Review

Paper Title	Authors	Contribution
Bloomberggpt: A large language model for finance[3]	Wu et al. (2023)	A specialized LLM, was introduced for finance industry having 50 billion parameters. It outperforms other LLMs in tasks related to general reasoning, benchmarks specific to the financial sector.
Fingpt: Open-source financial large language models[4]	Yang et al. (2023)	A financial language model which integrates diverse data sources like social media, financial filings, trends, and academic datasets for financial analysis insights.
Can gpt models be financial analysts? an evaluation of chatgpt and gpt-4 on mock cfa exams[5]	Callanan et al. (2023)	The study explored LLMs' potential in responding to CFA exam queries, suggesting their adaptability beyond traditional tasks.

Literature Review (Cont’d)

Paper Title	Authors	Contribution
Evaluating open-domain question answering in the era of large language models[6]	Kamalloo et al. (2023)	Human evaluation was proposed over lexical matching for precise large language model assessment in open-domain question answering.
Balancing specialized and general skills in llms: The impact of modern tuning and data strategy[7]	Zhang et al. (2023)	A thoroughly evaluation framework was constructed consisting of 45 tailored questions to evaluate performance of LLMs across various aspects including reliability, coherence, and business applicability where Llama-2 excelled in reasoning abilities.
FINANCEBENCH: A New Benchmark for Financial Question Answering[8]	Islam et al. (2023)	The Finance-bench dataset was introduced and benchmarked the performance of various LLMs on different experimental setup like single and shared vector store, prompt order etc. The study highlighted the limitations of LLMs in financial contexts, such as: uncertainty in right answers, a large number of wrong response and answer generation.

Literature Review (Cont’d)

Paper Title	Author s	Contribution	Major Limitation	Future Work
Enhancing Large Language Model Performance To Answer Questions and Extract Information More Accurately[9]	Zhang et al. (2024)	The study investigated enhancing the capabilities of large language models (LLMs) for financial question answering via few-shot learning and fine-tuning. GPT-3.5 Turbo, GPT4All, Llama2, and Claude was evaluated on Finance-Bench dataset. The significance of supervised fine-tuning and prompting techniques was highlighted for accuracy improvement.	• Lack of detailed elaboration on evaluation measures. • Absence of human evaluation for assessing model outputs. • Insufficient explanation of prompt development and workflow process. • Using of different methods employed for different models which introduce inconsistencies.	• Integrating human evaluation into fine-tuning process could enhance trustworthiness. • Standardizing prompt creation and fine-tuning procedures across models is needed. • Exploration of diverse model parameters and strategies for finance-centric domain is necessary.

Workflow of Research

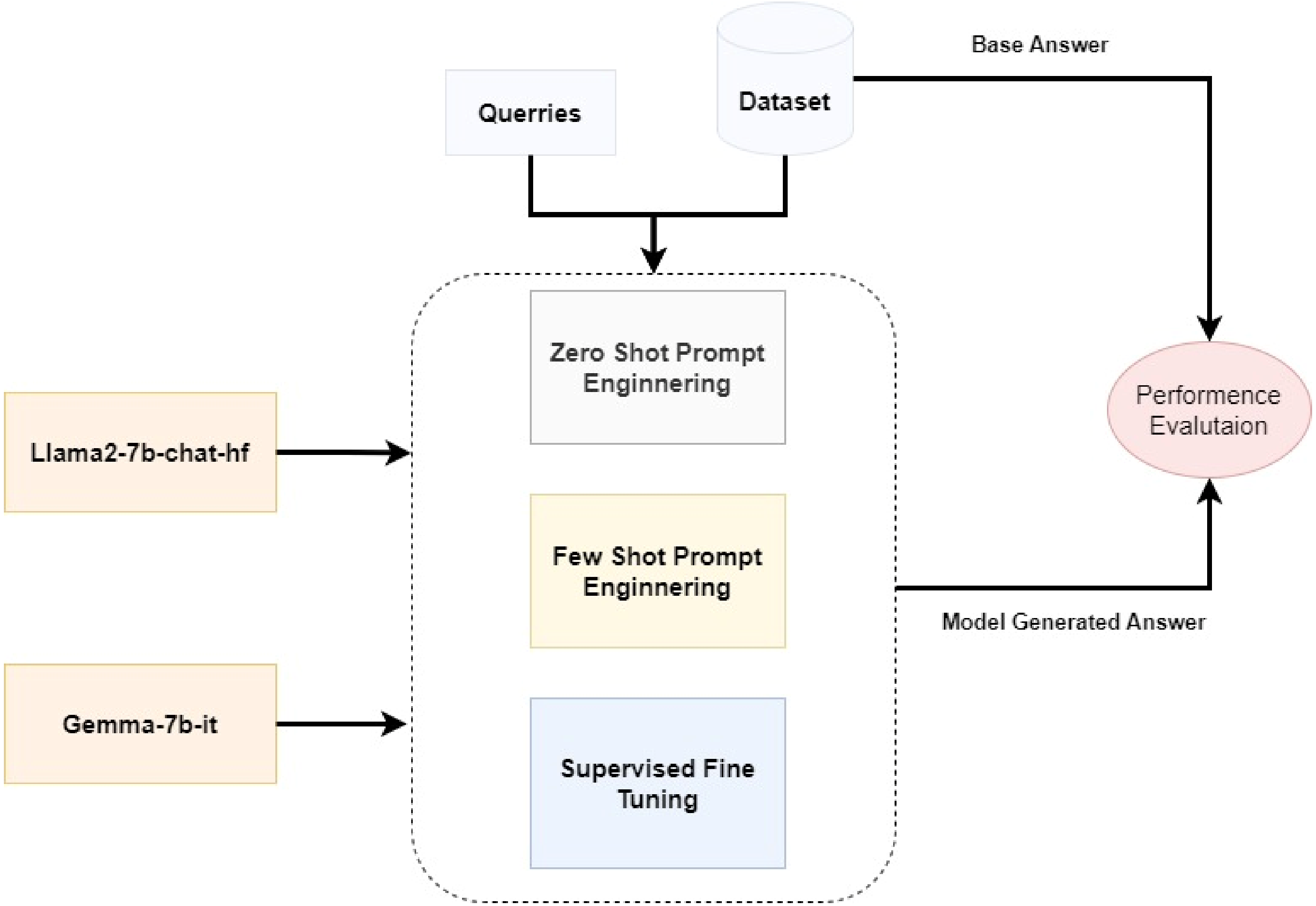

Fig 1: Workflow of Research

Dataset Description

FINANCE-BENCH

- A Dataset with 10,231 cases used for financial question answering tasks.
- To Test large language models' ability to answer financial questions.
- Emphasizes the need for improved evaluation methods for financial tasks.
- Study focused on the 150 rows of the main dataset.

Dataset Description (Cont'd)

Table 3: Samples from Finance-Bench dataset [10]

Question	Evidence_Text	Answer
Was there any drop in Cash & Cash equivalents between FY 2023 and Q2 of FY2024?	July 29, 2023 January 28, 2023 July 30, 2022 Cash and cash equivalents $ 1,093 $ 1,874 $ 840	Yes, there was a decline of ~42% between FY2023 and Q2 of FY 2024.
What drove operating margin change as of the FY22 for AMD? If operating margin is not a useful metric for a company like this, then please state that and explain why.	Operating income for 2022 was $1.3 billion compared to operating income of $3.6 billion for 2021. The decrease in operating income was primarily driven by amortization of intangible assets associated with the Xilinx acquisition.	The decrease in AMD's operating income was primarily driven by amortization of intangible assets associated with the Xilinx acquisition
At the Pepsico AGM held on May 3, 2023, what was the outcome of the shareholder vote on the shareholder proposal for a congruency report by Pepsico on net-zero emissions policies?	The shareholder proposal regarding a congruency report on net-zero emissions policies was defeated: For 19,718,780 Against 977,228,788	The shareholder proposal for a congruency report by Pepsico on net-zero emissions policies was defeated.

Dataset Description (Cont’d)

Table 4: Question difficulty categorizations for Test Set [11]

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

Methodology

Three methods were utilized on the Finance-Bench dataset:

- Zero shot prompt engineering
- Few-shot prompt engineering
- Supervised fine-tuning

Methodology (Cont'd)

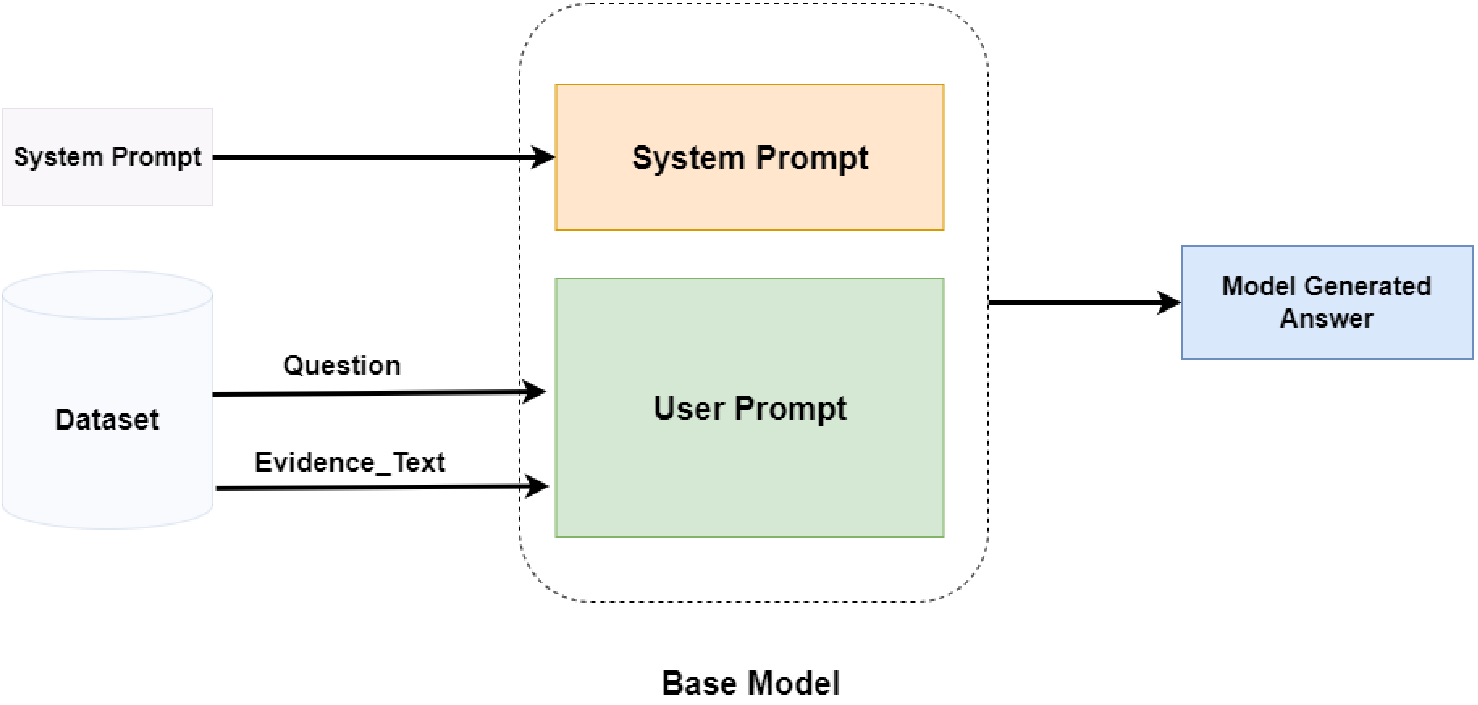

Fig 2: Zero Shot Prompt Engineering

Methodology (Cont’d)

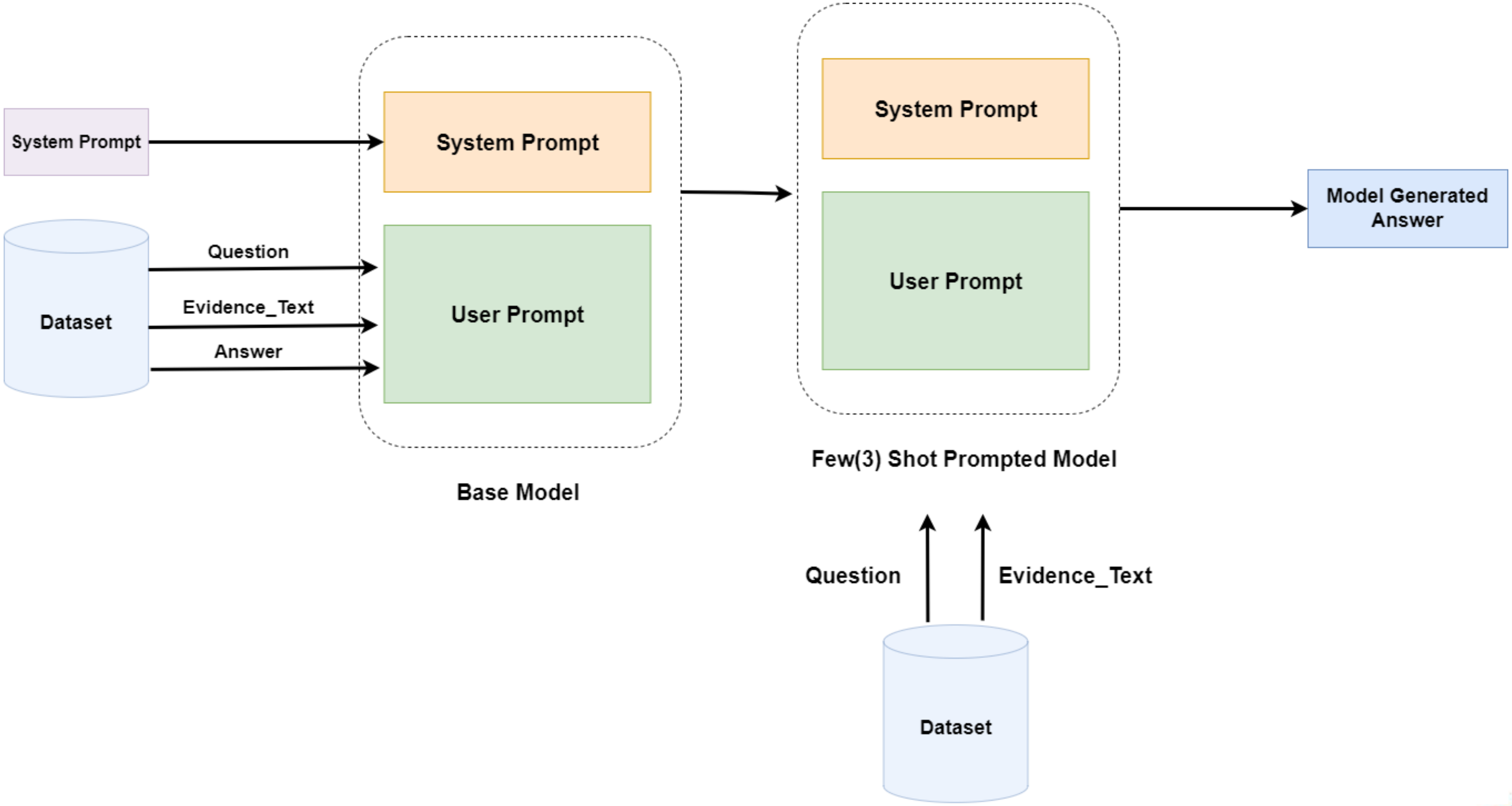

Fig 3: Few(3) Shot Prompt Engineering

Methodology (Cont'd)

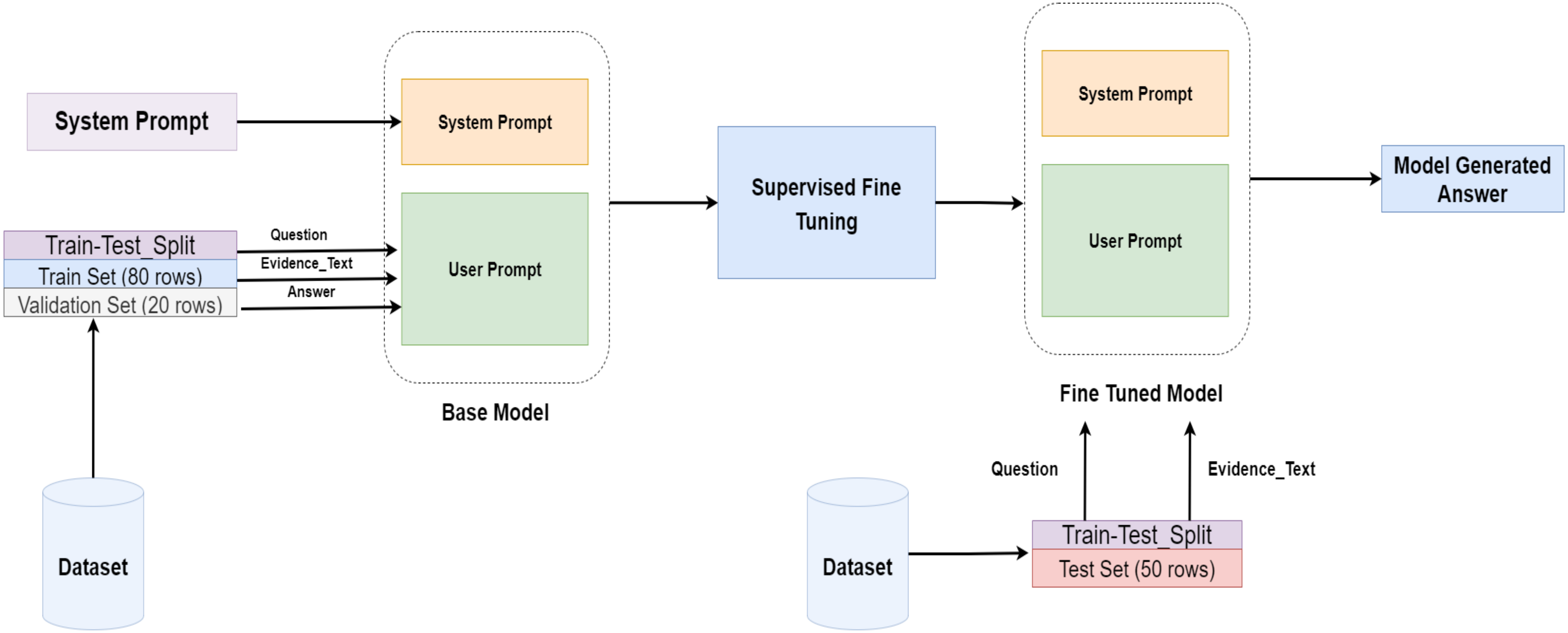

Fig 4: Supervised Fine Tuning

Methodology (Cont’d)

Table 5: Hyperparameters used for Llama2 and Gemma model for supervised fine-tuning

Hyperameters	Quantity
lora_config (r)	16
lora_alpha	64
lora_dropout	0.1
train_batch_size	4
optimizer	paged_adamw_32bit
learning_rate	2e-5
num_train_epochs	5

Results

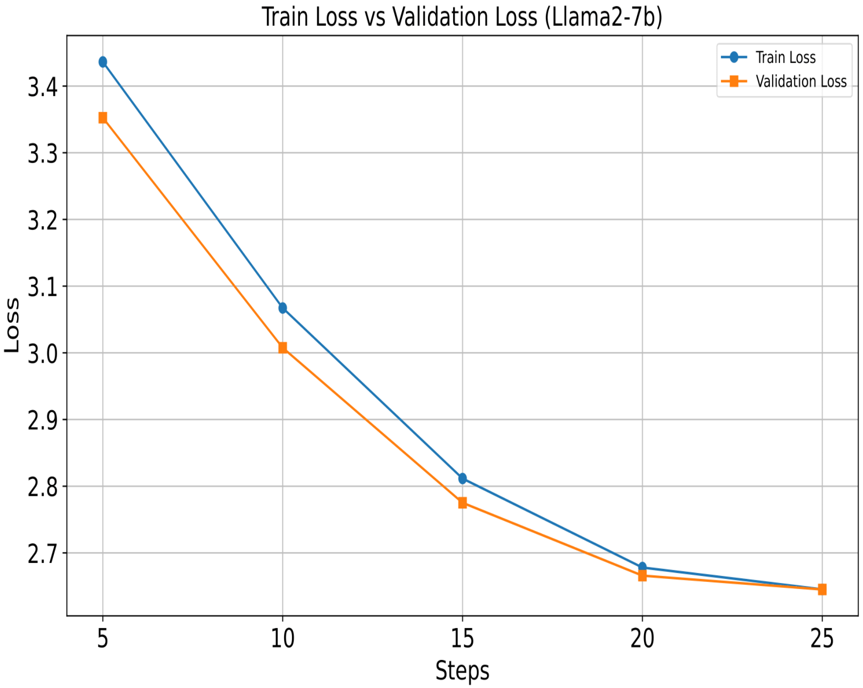

Fig. 5: Train loss vs Validation loss curve for Llama2

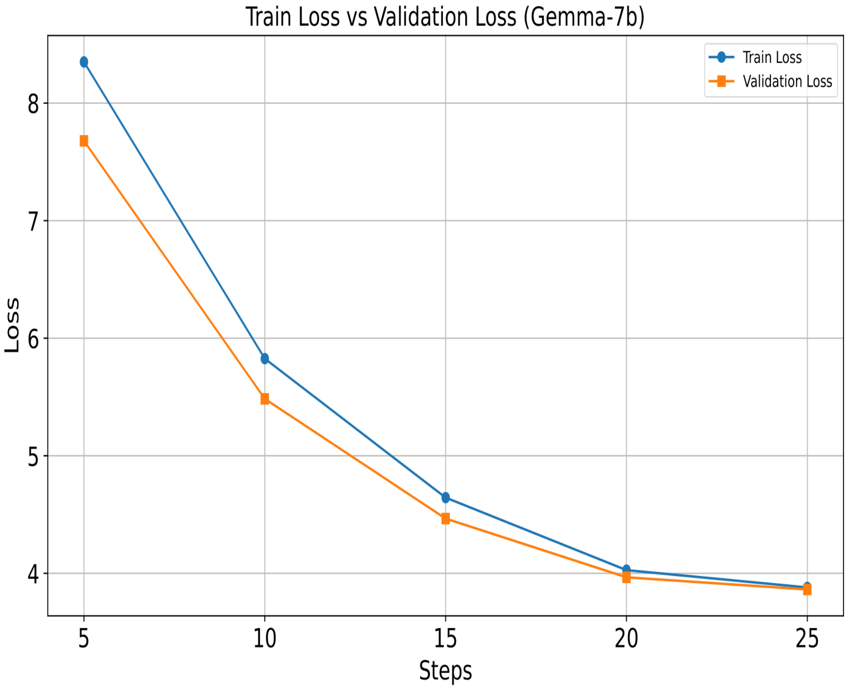

Fig. 6: Train loss vs Validation loss curve for Gemma

Results (Cont’d)

Evaluation Metrics

- ROUGE-L score : Evaluates text overlap by emphasizing the longest common subsequences (LCS) between generated and reference text.

- Cosine similarity: Evaluates the alignment of two vectors in multi-dimensional space by determining the cosine of the angle between them, enabling measurement of semantic similarity between texts.

- Human Evaluation: Ensures the factual correctness of generated responses, enhancing their relevance and reliability in decision-making processes.

Results (Cont’d)

Table 6: Total average ROUGE-L & Cosine Similarity score

Model	Technique	Llama2	Gemma
ROUGE-L	Zero Shot	**0.2463**	0.2125
	Few(3) Shot	**0.1885**	0.1596
	Supervised Fine Tuning	**0.2727**	0.1933
Cosine Similarity	Zero shot	0.1505	**0.1648**
	Few(3) Shot	0.1497	**0.1894**
	Supervised Fine Tuning	0.1673	**0.1942**

Results (Cont’d)

Table 7: Performance of the models on Human Evaluation

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

Results (Cont’d)

Table 8: Correctness Evaluation for Different Question Difficulties for Supervised Fine Tuning

Question Difficulty	Category	Llama2	Gemma
Easy	0-RETRIEVE	**61%**	7%
	1-COMPARE	**66%**	0%
	2-CALC-CHANGE	**14%**	0%
Hard	3-CALC-COMPLEX	**33%**	20%
	4-CALC-AND-JUDGE	**20%**	**20%**
	5-EXPLAIN FACTORS	50%	**100%**
	6-OTHER-ADVANCED	**50%**	**0%**

Conclusion

- Evaluation of Llama2 and Gemma model on financial question-answering tasks.
- Multi-phase methodology including zero-shot prompt engineering, few(3)-shot prompt engineering, and supervised fine-tuning was used.
- Llama2 outperforms Gemma in generating correct answers but produces more incorrect ones.
- Gemma minimizes incorrect answers but struggles with a higher rate of unanswered questions.
- Continued research are needed to enhance efficiency and accuracy of large language models in financial tasks prioritizing refining models to balance accuracy and completeness.

Future Works

- Incorporate more parameter-efficient fine-tuning approaches, such as prefix tuning, IA3 etc.
- Explore other training methods like Reinforcement Learning with Human Feedback (RLHF) and Unsupervised Learning.
- Utilize retrieval-augmented generation (RAG) technique for improved responses.
- Investigate performance gap between zero-shot and few-shot learning techniques.
- Include more diverse set of datasets like FinQA and ConvFinQA to ensure generalizability.
- Include more models like Claude, GPT, Falcon etc. for broader comparison.

References

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

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

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

[4]. Yang, Hongyang, Xiao-Yang Liu, and Christina Dan Wang. "Fingpt: Open-source financial large language models." *arXiv preprint arXiv:2306.06031* (2023).

[5]. 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

[6]. 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.

References (Cont’d)

[7]. Zhang, Zheng, Chen Zheng, Da Tang, Ke Sun, Yukun Ma, Yingtong Bu, Xun Zhou, and Liang Zhao. "Balancing specialized and general skills in llms: The impact of modern tuning and data strategy." *arXiv preprint arXiv:2310.04945* (2023).

[8]. 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.

[9]. 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.

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

[11]. Zooey Nguyen, Anthony Annunziata, Vinh Luong, Sang Dinh, Quynh Le, Anh HaiHa, Chanh Le, Hong An Phan, Shruti Raghavan, and Christopher Nguyen. 2024.Enhancing Q&A with Domain-Specific Fine-Tuning and Iterative Reasoning: AComparative Study. arXiv preprint arXiv:2404.11792 (2024). https://doi.org/10.48550/arXiv.2404.11792

Thank You