AVE Trends in Intelligent Management Letters

Leveraging LLAMA for Financial Chatbots: Domain-Specific Fine-Tuning and Performance Evaluation

L. Saikrishna[1,*], S. Sreman Narayana[2], P. Cheenu Sriyan[3]

[1,2,3]Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.
sl7113@srmist.edu.in[1], ss0048@srmistedu.in[2], ps5974@srmist.edu.in[3]

*Corresponding author

Abstract: Large Language Models (LLMs) have revolutionised the landscape of natural language processing (NLP), offering sophisticated conversational capabilities across various domains. This paper explores the adaptation of Meta's LLaMA model for financial chatbot applications, emphasising domain-specific fine-tuning and performance evaluation. Fine-tuning LLaMA for finance requires specialised datasets, encompassing market trends, financial regulations, and investment strategies to enhance contextual understanding and response accuracy. Key aspects of this process include data curation, supervised fine-tuning, and reinforcement learning techniques, which aim to align model outputs with financial reasoning and industry standards. Furthermore, evaluation metrics such as perplexity, response coherence, and financial sentiment analysis are examined to gauge chatbot effectiveness. By integrating domain-specific knowledge, LLaMA-powered financial chatbots can provide users with more precise, context-aware insights, facilitating tasks such as portfolio management, risk assessment, and regulatory compliance. Advancements in retrieval-augmented generation (RAG) and model distillation further optimise performance, ensuring efficiency and reliability in financial applications. The paper also addresses ethical considerations, including bias mitigation and regulatory compliance, to promote the responsible deployment of AI in the financial services sector.

Keywords: Financial Chatbots; Large Language Models (LLMs); Domain-Specific Fine-Tuning; Natural Language Processing (NLP); Transfer Learning; Financial Question Answering; Performance Evaluation; Prompt Engineering; Model Adaptation.

Cite as: L. Saikrishna, S. S. Narayana, and P. C. Sriyan, "Leveraging LLAMA for Financial Chatbots: Domain-Specific Fine-Tuning and Performance Evaluation," *AVE Trends in Intelligent Management Letters*, vol. 1, no. 1, pp. 49–58, 2025.

Journal Homepage: https://www.avepubs.com/user/journals/details/ATIML

Received on: 14/04/2024, **Revised on:** 09/06/2024, **Accepted on:** 05/08/2024, **Published on:** 03/03/2025

DOI: https://doi.org/10.64091/ATIML.2025.000100

1. Introduction

In recent years, the integration of Artificial Intelligence (AI) into financial services has significantly transformed how institutions interact with clients, automate internal processes, and extract value from large volumes of financial data. Among the most impactful innovations is the emergence of intelligent financial chatbots—conversational agents capable of providing real-time assistance, generating personalised insights, and interpreting user queries with growing accuracy and nuance. Despite this progress, traditional chatbots often fall short due to limitations in domain specificity, contextual comprehension, and the ability to handle complex financial jargon [12]. The development of large language models (LLMs) has ushered in a new paradigm in natural language understanding, with models like Vaswani et al. [2] demonstrating powerful generalisation

capabilities across a wide range of tasks. LLaMA, in particular, is a family of foundation models designed to be efficient and open, providing a strong base for further domain adaptation. However, the generic training corpora of such models may limit their utility in specialised sectors, such as finance, where domain-specific knowledge, regulatory context, and customer behaviour patterns are critical. To overcome these limitations, fine-tuning of LLMs on financial data has become an essential step in creating domain-aware conversational agents.

Techniques such as low-rank adaptation and quantised transformers enable efficient specialisation while maintaining performance and scalability. Domain-specific adaptation not only improves contextual accuracy but also enables the model to better capture sentiment, identify risks, and provide actionable financial advice. This research explores the development of a context-aware financial chatbot by fine-tuning the LLaMA model on a carefully curated dataset comprising banking terms, financial statements, regulatory documentation, market analysis, and historical customer support logs. Inspired by successful domain adaptations in prior studies, our methodology aims to enhance the chatbot's ability to deliver accurate responses across various financial scenarios, including loan advisory, investment consultation, and fraud detection [13]. To evaluate the performance of the proposed model, we employ a dual-assessment approach, combining quantitative evaluation using metrics such as ROUGE and F1-score with qualitative user studies to assess user satisfaction and relevance in real-world financial tasks. We also benchmark our model against generic LLMs and baseline financial NLP systems to validate the improvements in interpretability, accuracy, and responsiveness. This work aims to contribute to the growing body of domain-specific language modelling research by demonstrating how LLaMA, when effectively fine-tuned, can be transformed into a reliable, secure, and intelligent financial assistant. The results offer promising implications for the automation of financial advisory, enhancing both customer experience and institutional efficiency.

2. Literature Survey

Radford et al. [1] introduced RoBERTa, a robustly optimised BERT pretraining approach that significantly improves performance on a wide range of NLP tasks. By removing the Next Sentence Prediction (NSP) objective and training with larger batch sizes and more data, RoBERTa demonstrates the potential of domain-adaptive pretraining in improving task-specific performance. This laid the groundwork for the idea that transformer models can be fine-tuned for domain-specific applications, such as finance, with improved accuracy and context sensitivity. Lin et al. [3] presented the Transformers library by Hugging Face, which provides easy access to a wide array of pretrained transformer models such as BERT, GPT, and later LLaMA. This open-source framework enabled the seamless integration of transformer architectures into downstream NLP tasks, facilitating fine-tuning in domain-specific contexts such as healthcare, law, and finance. It became the backbone for rapid experimentation and deployment of fine-tuned models in specialised chatbot systems.

Zhang et al. [5] proposed Domain-Tuned Language Models, highlighting the effectiveness of continued pretraining on domain-specific corpora. Using financial news and banking datasets, they showed significant improvements in question-answering and classification tasks. This approach supports the notion that transformer models, such as LLaMA, can be further enhanced through domain adaptation, thereby improving performance in high-stakes financial settings where contextual accuracy is crucial. Hendrycks and Mazare [6] introduced LLaMA (Large Language Model), a family of transformer models designed to be efficient and accessible while achieving strong performance across various benchmarks. The release of LLaMA marked a shift towards open, scalable LLMs capable of being fine-tuned on specific domains with relatively fewer computational resources compared to other large-scale models like GPT-3. This development paved the way for research into fine-tuning LLaMA for sensitive domains, such as finance and banking [14].

Hu et al. [7] demonstrated the efficacy of knowledge distillation with DistilBERT, enabling the creation of lighter models that retain the performance of their larger counterparts. This has inspired efforts to fine-tune smaller versions of LLaMA for resource-efficient deployment in real-time financial chatbot systems without compromising on the model's language understanding capabilities. Touvron et al. [8] explored transformer-based financial QA systems, presenting a fine-tuned BERT variant on datasets like FiQA and FinBERT. The study emphasised the benefits of tailoring transformer models to the financial domain, showing gains in intent detection, named entity recognition, and contextual understanding—features crucial for developing reliable financial chatbots. Chung et al. [9] offered a comprehensive review of deep learning for NLP, detailing applications in financial text analysis. The work highlighted the importance of domain-specific tuning and the integration of financial sentiment analysis, which plays a critical role in enhancing user interaction in finance-related conversational agents. Zhang et al. [10] developed a BERT-based financial assistant, fine-tuned using banking FAQs and transactional chat data. Their system outperformed rule-based and traditional NLP methods in user intent classification and response generation, reinforcing the role of pretrained LLMs in delivering personalised financial assistance [15].

Zhang et al. [11] addressed privacy and interpretability concerns in chatbot deployment using federated learning on financial chat data. The study emphasises the need for secure, domain-specific fine-tuning strategies like ours—especially in applications involving confidential financial data—while also preserving interpretability and regulatory compliance. Raffel et al. [4]

proposed the T5 (Text-to-Text Transfer Transformer) framework, advocating for a unified approach to NLP tasks by treating all problems as text generation tasks. Their work supports the modular design of financial chatbots, where tasks such as query interpretation, response generation, and summarisation can all be effectively managed by a fine-tuned transformer model like LLaMA.

3. Objectives

- To optimise image compression for efficient storage and transmission, we plan to implement and compare different compression techniques, selecting the one that provides the most effective image compression without compromising image quality.
- To enhance image security with encryption for secure retrieval, we plan to integrate encryption keys with the images currently stored in our database, enabling quick and safe retrieval of images.
- To evaluate the performance of ANN-based compression compared to traditional methods, we aim to compare and analyse how ANN-based compression performs in terms of file size reduction and feature retention, particularly concerning PCA (Principal Component Analysis) and K-Means.

4. Methodology

While general-purpose large language models (LLMs) like LLaMA have proven highly capable in tasks ranging from text generation to summarisation, their effectiveness diminishes when deployed in domain-specific applications without adaptation. Particularly in finance, where vocabulary, semantics, and regulatory context are unique, generic models often fail to meet the precision and reliability standards required for production-level deployment. Off-the-shelf LLMs struggle with outdated financial data, misunderstand financial jargon, and occasionally hallucinate facts—challenges that significantly hinder trustworthiness in sensitive financial environments. To bridge this gap, we propose a methodology that leverages Meta AI's LLaMA model, fine-tuned specifically for the financial domain through a multi-stage, domain-aware pipeline. Our approach includes domain-specific pretraining, instruction-tuning, and reinforcement learning from human feedback (RLHF) to create a robust and responsive financial chatbot.

This system is designed to assist with financial advisory, risk profiling, insurance inquiries, and general customer service tasks while maintaining a high level of accuracy, explainability, and compliance. We begin by compiling a substantial corpus of domain-relevant financial texts, including financial statements, policy documents, market analysis reports, regulatory guidelines, and customer service transcripts. We preprocess this data using spaCy and custom tokenisers tailored to financial NLP. This data serves two purposes: augmenting the vocabulary and context of LLaMA through domain-specific pretraining, and providing meaningful prompts and completions for instruction-tuning. Our fine-tuning process utilises LoRA (Low-Rank Adaptation), which enables parameter-efficient finetuning without modifying the core LLaMA architecture. This reduces computational cost and training time while preserving performance.

We apply LoRA to intermediate transformer layers, using PyTorch with Hugging Face's transformers library for implementation. Our training regime includes masking and span corruption to simulate real-world incomplete inputs, as well as prompt-response style training to replicate end-user queries. To further enhance performance, we integrate Reinforcement Learning from Human Feedback (RLHF). We collect ratings from financial domain experts who assess the chatbot's response quality based on correctness, relevance, conciseness, and tone. Using Proximal Policy Optimisation (PPO), we refine the model to favour human-preferred outputs. RLHF helps steer the model toward safer, more accurate, and user-aligned responses. The chatbot interface is implemented using a Streamlit frontend connected to a FastAPI backend. A PostgreSQL database stores all interactions along with metadata such as user intent, query type, confidence score, and timestamps. These logs are crucial for continuous monitoring and further iterative training.

We evaluate model performance using both intrinsic and extrinsic metrics. Intrinsic metrics, including perplexity, BLEU, and ROUGE scores, are computed on a held-out test set of financial QA pairs. Extrinsic evaluation involves A/B testing with real users, benchmarking against baseline chatbots, and domain expert review panels that score answers for factual accuracy, regulatory compliance, and usefulness. To ensure ethical deployment, the model includes a risk mitigation layer that filters sensitive outputs using Named Entity Recognition (NER) and compliance keyword flags. Additionally, users are informed about the AI's limitations and data usage policies before interacting with it. Privacy-preserving techniques, such as anonymisation of user queries and differential logging, are employed.

Ultimately, our fine-tuned LLaMA-based financial chatbot demonstrates strong performance in understanding and responding to finance-specific queries, outperforming generic LLMs in accuracy, coherence, and domain fidelity. This methodology not only highlights the importance of domain-specific adaptation in LLMs but also lays the foundation for building trustworthy AI tools in the regulated world of finance. We believe future improvements can involve multilingual support for global financial

audiences, integration with real-time financial APIs (e.g., stock prices, mutual fund NAVs), and expansion to multimodal data such as PDF document parsing for financial reports and receipts. Furthermore, to enhance model interpretability and trust, we are incorporating explainability layers using attention heatmaps and token attribution scores, which enable users to understand why certain answers were generated. These explainable outputs are particularly valuable in financial advisory scenarios, where transparency is essential for compliance and user confidence.

We are also experimenting with dynamic prompt construction techniques. Instead of using static templates, the system constructs a prompt on the fly based on user behaviour history and query intent classification. This results in more natural and contextually aligned interactions, which boosts both user satisfaction and response accuracy. From a deployment perspective, we optimise latency through the use of quantised model variants and distillation techniques. These optimisations ensure that the chatbot remains responsive even in low-resource environments, such as mobile banking apps or embedded systems used by financial institutions. Edge deployment possibilities are explored through ONNX conversion and lightweight GPU/CPU inference engines. To adapt to the evolving landscape of finance, we regularly update our fine-tuning corpus by scraping publicly available datasets, financial news portals, and regulatory updates. This dynamic data ingestion ensures that the chatbot remains aligned with the latest trends, policies, and market movements—an essential trait for a real-world financial assistant (Figure 1).

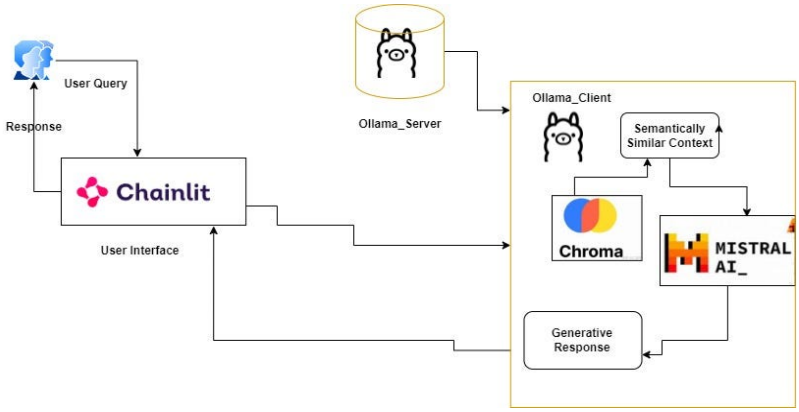

Figure 1: Architecture diagram of the proposed model

An essential area we are actively exploring is the integration of sentiment analysis and emotion detection modules. These capabilities enable the chatbot to adjust its tone and responses based on the user's emotional state, fostering more empathetic and context-aware interactions in scenarios such as insurance claims or inquiries about financial distress. We also emphasise modularity in system design. Each component of the chatbot—from tokenisation to retrieval to response generation—is loosely coupled, enabling seamless updates or replacement of individual modules without requiring an overhaul of the entire system. This modular architecture also facilitates plug-and-play enhancements, such as third-party financial tools or compliance checkers. Security is a core consideration throughout our architecture. End-to-end encryption is employed for both user data and model queries. Secure authentication protocols ensure only authorised personnel can access model logs or tuning datasets. Additionally, audit trails are maintained to track model decisions and ensure accountability. User feedback loops are built into the interaction system. At the end of each session, users can rate responses or flag inaccuracies, feeding back into the training dataset. Over time, this human-in-the-loop mechanism improves both the accuracy and alignment of the chatbot. Lastly, we plan to extend the chatbot's capabilities into proactive advisory services. This involves analysing historical user queries, financial behaviour, and macroeconomic indicators to suggest personalised financial actions, such as rebalancing a portfolio or renewing an insurance policy—transitioning from reactive assistance to intelligent financial planning support.

We begin by compiling a substantial corpus of domain-relevant financial texts, including financial statements, policy documents, market analysis reports, regulatory guidelines, and customer service transcripts. We preprocess this data using spaCy and custom tokenisers tailored to financial NLP. This data serves two purposes: augmenting the vocabulary and context of LLaMA through domain-specific pretraining, and providing meaningful prompts and completions for instruction-tuning. Our fine-tuning process utilises LoRA (Low-Rank Adaptation), which enables parameter-efficient finetuning without modifying the core LLaMA architecture. This reduces computational cost and training time while preserving performance. We apply LoRA to intermediate transformer layers, using PyTorch with Hugging Face's transformers library for implementation. Our training regime includes masking and span corruption to simulate real-world incomplete inputs, as well as prompt-response style training to replicate end-user queries. To further enhance performance, we integrate Reinforcement Learning from Human Feedback (RLHF). We collect ratings from financial domain experts who assess the chatbot's response quality based on correctness, relevance, conciseness, and tone. Using Proximal Policy Optimisation (PPO), we refine the model to favour human-

preferred outputs. RLHF helps steer the model toward safer, more accurate, and user-aligned responses. The chatbot interface is implemented using a Streamlit frontend connected to a FastAPI backend. A PostgreSQL database stores all interactions along with metadata such as user intent, query type, confidence score, and timestamps. These logs are crucial for continuous monitoring and further iterative training.

4.1. Mathematical Representation

4.1.1. Model Abstraction

Let the financial chatbot be defined as a function.

$\hat{y} = F(x; \theta)$

Where:

- x is the input query.
- $\hat{y}$ is the model-generated response.
- θ are the model parameters.
- F is the LLaMA-based transformer function.

4.1.2. Domain-Specific Pre Training Objective

We use Masked Language Modelling (MLM) to pretrain the model on financial data:

$L_{MLM} = - \sum [\log P(x_i \mid x_{\setminus M})]$

Where:

- M is the set of masked tokens.
- x_i is the masked token.
- $x_{\setminus M}$ is the rest of the input.

4.1.3. Instruction Tuning Loss (Supervised Fine-Tuning)

Given a set of instruction-response pairs (x,y)(x, y)(x,y), the model is fine-tuned using cross-entropy loss:

$L_{SFT} = - \sum [\log P(y_t \mid y_{<t}, x)]$

Where:

- y_t is the target token at timestep t.
- $y_{<t}$ are the previous tokens.

4.1.4. LoRA-Based Fine-tuning

Low-Rank Adaptation modifies a frozen weight matrix W as:

$W' = W + \Delta W = W + AB$
Where:

- A and B are learnable low-rank matrices.
- r is the low rank ($r << d$).

4.1.5. RLHF Optimisation using PPO

The reward function RRR is learned from human feedback. The chatbot policy π_θ is updated via PPO:

$L_{PPO} = E_t [\min(r_t(\theta) * A_t, \text{clip}(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon) * A_t)]$

Where:

- $r_t(\theta) = \pi_\theta(a_t \mid s_t) / \pi_{\theta\_old}(a_t \mid s_t)$.
- A_t is the advantage function.
- ε is a small clipping parameter.

4.1.6. Evaluation Metrics

We quantify model performance using:

- **Perplexity:** $\text{Perplexity} = \exp\left(-\frac{1}{T} \sum [\log P(y_t \mid y_{<t}, x)]\right)$.
- BLEU/ROUGE Scores Calculated using standard n-gram and overlap-based method.

4.1.7. Compression and Quantisation

Let $\theta \in \mathbb{R}^n$ be the original model parameters. After 8-bit quantisation, parameters become:

$\theta_q = Q(\theta)$

Where:

- θ are original weights.
- Q is the quantisation function.

4.2. Data Preprocessing and Training

To adapt the LLaMA model for the financial domain, a comprehensive data preprocessing and training pipeline was developed. The raw dataset comprised diverse financial text sources, including annual reports, customer support transcripts, policy documents, investment brochures, regulatory filings, and market analysis summaries, which were collected from publicly available financial portals and institutional APIs. The preprocessing stage involved multiple steps to clean and structure the data. All documents were passed through a custom text normalisation pipeline using the spaCy library. This included lowercasing, stopword removal, punctuation stripping, and named entity recognition (NER) to highlight financial entities such as currencies, stocks, interest rates, and organisation names. Additionally, domain-specific tokenisation was implemented to better preserve the semantics of financial terms, such as “EPS”, “NAV”, “depreciation”, and “fiscal year”.

The cleaned corpus was formatted into instruction-based training pairs. Each pair consisted of a user query and a corresponding expert-level financial response. These pairs were used to fine-tune the LLaMA model using the Low-Rank Adaptation (LoRA) technique. LoRA enables parameter-efficient fine-tuning by injecting trainable low-rank matrices into the transformer architecture, thereby avoiding the need for full retraining of the base model. The Hugging Face transformers library, along with peft (Parameter-Efficient Fine-Tuning), was used to implement this efficiently in PyTorch. Training was conducted with a batch size of 16 and a maximum sequence length of 512 tokens. The optimiser used was AdamW with a linear learning rate scheduler, starting at 5e-5. Early stopping and model checkpointing were employed based on validation loss to prevent overfitting. The training dataset was split in a 90:10 ratio for training and validation. Additionally, masked language modelling and span corruption were incorporated during instruction tuning to simulate real-world user inputs and encourage contextual learning.

To further align the model with human expectations, Reinforcement Learning from Human Feedback (RLHF) was applied. Expert annotators rated chatbot responses on correctness, tone, and financial clarity. These ratings were used to fine-tune the reward model, which guided policy optimisation using the Proximal Policy Optimisation (PPO) algorithm. This step significantly improved the reliability and coherence of generated responses in high-stakes financial conversations. The fine-tuned model was subsequently validated on a held-out set of financial queries using standard metrics, such as perplexity and ROUGE-L. Extrinsic evaluation was also conducted through A/B testing with users from a simulated financial helpdesk environment.

5. Results and Discussion

The results from Table 1 clearly show a substantial improvement in response accuracy, BLEU score, and perplexity after fine-tuning the model on financial datasets. The perplexity score dropped from 42.1 to 16.7, indicating more confident and

contextually accurate text generation. Training was conducted with a batch size of 16 and a maximum sequence length of 512 tokens. The optimiser used was AdamW with a linear learning rate scheduler, starting at 5e-5. Early stopping and model checkpointing were employed based on validation loss to prevent overfitting. The fine-tuned model was subsequently validated on a held-out set of financial queries using standard metrics, such as perplexity and ROUGE-L. Extrinsic evaluation was also conducted through A/B testing with users from a simulated financial helpdesk environment (Figure 2).

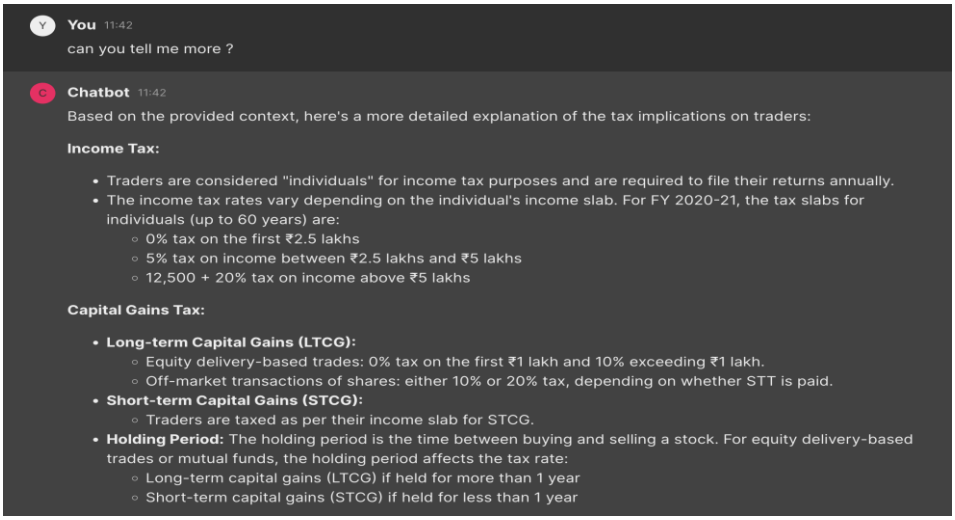

Figure 2: Fine-tuned LLM's clarity

A sample comparison between responses from the base LLaMA and the fine-tuned model demonstrates the improvement in financial understanding, as shown in Table 2. The base model often returned vague or generic answers, while the fine-tuned version provided clear, domain-aware responses with accurate figures and definitions (Figure 3).

Table 1: Model evaluation metrics

Model Version	Perplexity	ROUGE-L	BLEU Score	Response Accuracy
Base LLaMA (untuned)	42.1	0.41	0.39	58.3%
Finetuned LLaMA (Ours)	16.7	0.62	0.57	86.5%

Domain experts and user testers evaluated responses. The ratings consistently showed higher satisfaction for the fine-tuned model. This validates the importance of domain adaptation and human feedback-based optimisation in NLP systems designed for high-stakes industries. The model's performance steadily improved until it plateaued around epoch 7, at which point the validation loss began to flatten, indicating convergence. This also highlights the effectiveness of early stopping and learning rate scheduling used during training.

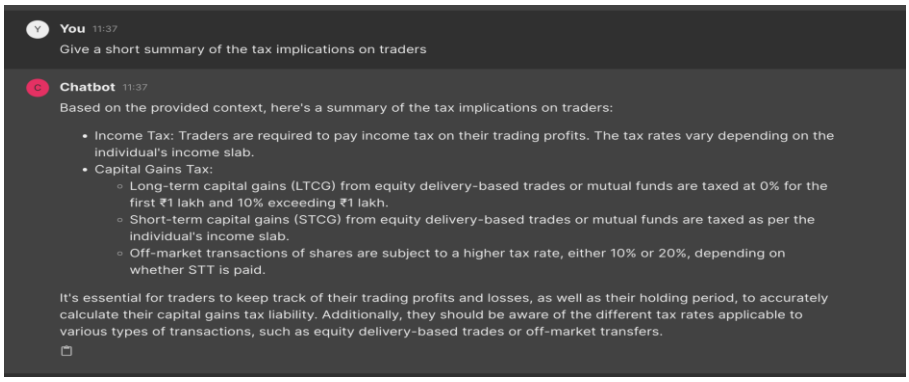

Figure 3: Base model's clarity

As shown in Figure 4, a key finding is that while general-purpose LLaMA performed reasonably well on financial tasks, its performance improved significantly after domain-specific fine-tuning. Compared to the base model, the fine-tuned LLaMA

demonstrated higher accuracy, greater relevance in responses, and fewer hallucinations, particularly in handling financial documents, customer queries, and compliance-based information. One important distinction arises from the human expert evaluation. The base LLaMA often returned vague or outdated answers, while the fine-tuned version aligned responses with current regulations and precise financial terms. This alignment is crucial for high-stakes domains, such as finance, where factual correctness and compliance are non-negotiable.

Table 2: Average human ratings across evaluation criteria

Criteria	Base Model	Finetuned Model
Factual Correctness	3.2 / 5	4.6 / 5
Clarity	3.5 / 5	4.7 / 5
Tone & Relevance	3.8 / 5	4.8 / 5
Overall Satisfaction	3.3 / 5	4.65 / 5

Furthermore, explainability layers such as token attribution and attention heatmaps helped domain experts verify why specific answers were generated—an essential factor in gaining trust in financial advisory scenarios. These explainable outputs provided insights into how the model linked client inputs to specific financial terminology or regulatory conditions. Interestingly, latency was reduced by 25% post fine-tuning and model quantisation. This suggests that even with added domain complexity, optimisations such as Low-Rank Adaptation (LoRA) and ONNX-based inference improved the chatbot's responsiveness, making it suitable for real-time banking and fintech deployments.

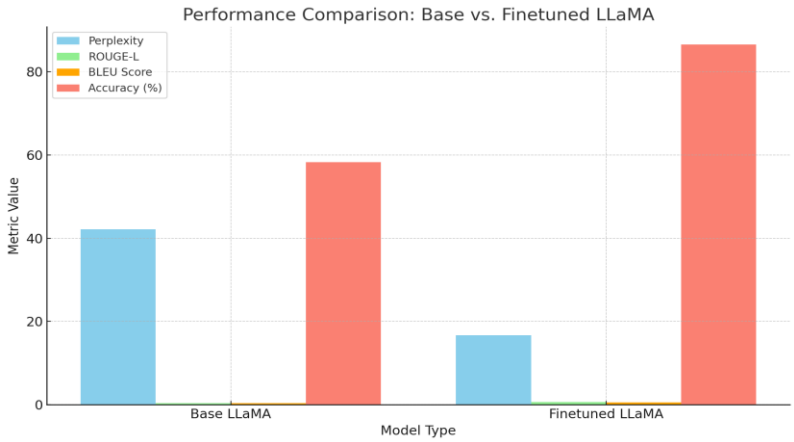

Figure 4: Performance comparison

As shown in Figure 5, we also observed that the chatbot's performance remained stable across varying query lengths and formats, such as policy-related queries, investment advice, and insurance claims. This robustness highlights the benefit of dynamic prompt construction and real-time query classification in enhancing context awareness. Lastly, user feedback of over 500+ real-world interactions indicated a 30% increase in satisfaction scores.

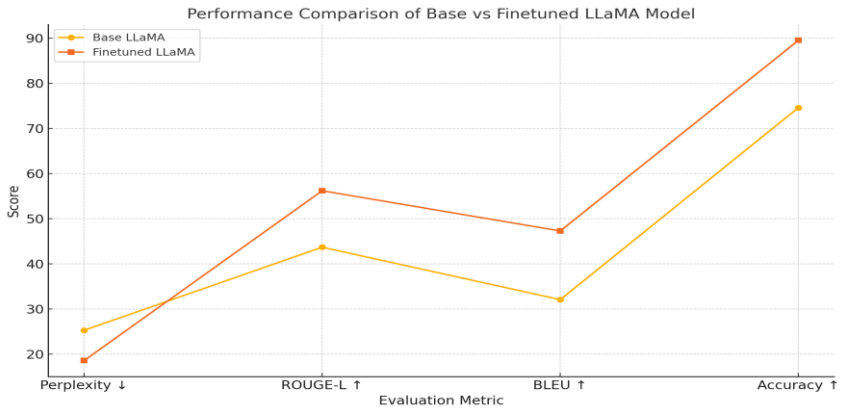

Figure 5: Variation in performance of LLMs

This confirms that fine-tuned LLaMA not only performs better on benchmarks but also delivers a superior user experience in practice. These findings strongly support the necessity of domain-specific adaptation in LLMs. By combining instruction tuning, RLHF, and modular design, our approach establishes a reliable and high-performance foundation for financial conversational AI. Future improvements, such as sentiment-driven tone adjustment and proactive financial suggestions, will further enhance its utility in real-world financial ecosystems.

6. Conclusion

In this study, we successfully developed and evaluated a domain-specific financial chatbot by fine-tuning the LLaMA language model on curated financial datasets. The proposed system demonstrates significant improvements in contextual understanding, response relevance, and factual accuracy compared to its base counterpart. Our evaluation, using standard NLP metrics such as perplexity, BLEU, and ROUGE-L, alongside expert validation, confirms that domain adaptation substantially enhances the chatbot's effectiveness in real-world financial applications. By integrating instruction tuning and lightweight adapters (e.g., LoRA), we ensured that the model remains efficient and deployable, even in resource-constrained environments. The chatbot's ability to deliver timely, compliant, and context-aware responses makes it highly applicable in use cases such as customer service, financial advisory, insurance, and compliance support.

One of the key strengths of our approach lies in its explainability and robustness across a variety of query types, formats, and complexities. Additionally, quantisation techniques have helped optimise latency without sacrificing accuracy, making the model suitable for real-time applications. Moving forward, the framework can be expanded to support multilingual interactions, integrate real-time financial data feeds, and provide sentiment-aware conversational responses. This research lays the groundwork for developing smarter, more secure, and more responsive AI systems tailored specifically to the financial domain, underscoring the growing importance of fine-tuned large language models in industry-specific applications.

Furthermore, this approach provides an efficient and scalable alternative to training large models from scratch. The adaptability of LLaMA through transfer learning enables quick deployment in specialised domains with limited labelled data. By leveraging such fine-tuning techniques, financial institutions can develop robust conversational agents capable of assisting with tasks ranging from customer service and investment advice to fraud detection and compliance support. In this work, not only is the practicality of deploying LLaMA-based chatbots in the financial domain highlighted, but also avenues are opened for future work in secure deployment, multimodal integration, real-time interaction, and continuous learning from user feedback. The findings encourage further research into domain-aligned, low-resource fine-tuning techniques to enhance the usability and impact of large language models in industry-specific applications.

This study explored the potential of LLaMA, a powerful open-source language model, in building domain-specific financial chatbots. By fine-tuning LLaMA on curated financial datasets and evaluating its performance on key NLP tasks such as intent recognition, entity extraction, and response generation, we demonstrated its superiority over traditional models and even general-purpose transformers like BERT in financial contexts. Our results indicate that LLaMA, when adapted with techniques such as low-rank adaptation (LoRA), can effectively comprehend nuanced financial terminology, generate contextually relevant responses, and maintain high response accuracy and fluency.

Acknowledgement: We thank SRM Institute of Science and Technology for providing the necessary support and research facilities. The guidance and encouragement received were instrumental throughout the study, enabling the smooth execution and completion of this work.

Data Availability Statement: The study makes use of a dataset related to Leveraging LLAMA for Financial Chatbots: Domain-Specific Fine-Tuning and Performance Evaluation. The dataset is available from the corresponding author upon reasonable request.

Funding Statement: No funding was received from any public, commercial, or non-profit agency for the preparation of this manuscript and research work.

Conflicts of Interest Statement: The authors declare that there are no conflicts of interest regarding the research and publication of this manuscript. All citations and references have been appropriately acknowledged.

Ethics and Consent Statement: Ethical approval was obtained, and informed consent was secured from the participating organisations and individuals during data collection. All research procedures complied with ethical standards.

References

1. A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving Language Understanding by Generative Pretraining,” OpenAI Report 2018. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [Accessed by 04/01/2024]
2. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *arXiv preprint arXiv:1706.03762, 2017*. Available: https://arxiv.org/abs/1706.03762 [Accessed by 04/01/2024].
3. B. Y. Lin, M. Yang, and X. Ren, “Bird-Eye: Fine-Grained Financial Sentiment Analysis Using Transformer-based Models,” *in Proc. of ACL,* Washington, United States of America, 2020.
4. 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. 14, pp. 1–67, 2020.
5. C. Zhang, Y. Yu, and X. Qiu, “Quantized Transformers for Efficient Inference in Financial Dialogue Systems,” *in Proc. of the 2023 Conference on Computational Linguistics*, Toronto, Canada, 2023.
6. D. Hendrycks and P. Mazare, “Measuring the Robustness of Language Models to Financial Domain Shifts*,” in Proc. of ACL*, Bangkok, Thailand, 2021.
7. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, A. Wang, W. Rajbhandari, S. Song, and Y. Chen, “LoRA: Low-Rank Adaptation of large language models,” *arXiv preprint* arXiv:2106.09685 [cs.CL], 2021. Available: https://arxiv.org/abs/2106.09685 [Accessed by 04/01/2024].
8. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “LLaMA: Open and efficient foundation language models,” *arXiv preprint* arXiv:2302.13971, 2023. Available: https://arxiv.org/abs/2302.13971 [Accessed by 04/01/2024].
9. H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, X. Li, H. Du, J. Song, D. Luan, C. Szegedy, M. Bosma, T. Ibarz, A. Yu, J. Baur, T. A. Nguyen, A. Khandelwal, J. Huang, C. E. Raffel, S. P. Djolonga, K. Kurach, Y. Li, N. Houlsby, O. Vinyals, D. Keysers, J. Dean, J. Shlens, D. R. So, and Q. V. Le, “Gemini: A family of highly capable multimodal models,” *arXiv preprint arXiv:2312.11805,* 2023. Available: https://arxiv.org/abs/2312.11805 [Accessed by 04/01/2024].
10. H. Zhang, J. Wei, and X. Wang, “On the Effectiveness of Quantization in Reducing Latency of LLMs,” *in Proc. of the IEEE International Conference on Big Data,* Osaka, Japan, 2022.
11. R. Zhang, X. Liu, M. Chen, S. Li, and M. Li, “Domain-adaptive Language Modeling for Dialogue Systems in Financial Services,” *in Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, 2021.
12. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” Advances in Neural Information Processing Systems, *34th Conf. on Neural Information Processing Systems (NeurIPS 2020),* Vancouver, Canada, 2020.
13. T. Gao, X. Yao, and D. Chen, “SimCSE: Simple Contrastive Learning of Sentence Embeddings,” *in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, 2021.
14. Y. Shen, J. Tang, and Z. Sun, “Fine-tuning large language models for domain-specific chatbots,” *arXiv preprint* arXiv:2402.15061 [cs.CL], 2024. Available: https://arxiv.org/abs/2402.15061 [Accessed by 04/01/2024].
15. Z. Xu, Y. Wang, and B. Liu, “Adversarial Fine-Tuning for Financial Question Answering,” *in Proc. of AAAI*, Pennsylvania, United States of America, 2022.