Developing a small language model for financial data

1st Kasuni Hansachapa
Department of Computer Engineering
University of Peradeniya
Peradeniya, Sri Lanka
e19131@eng.pdn.ac.lk

2nd Dr.Asitha Bandaranayake
Department of Computer Engineering
University of Peradeniya
Peradeniya, Sri Lanka
asithab@eng.pdn.ac.lk

3rd Prof.Roshan G. Ragel
Department of Computer Engineering
University of Peradeniya
Peradeniya, Sri Lanka
roshanr@eng.pdn.ac.lk

***Abstract*—This literature review examines the development and optimization of Small Language Models (SLMs) for financial applications as a cost-effective alternative to Large Language Models (LLMs). The study explores various techniques such as quantization, pruning, QLoRA fine-tuning, and knowledge distillation to enhance efficiency while mitigating computational constraints. Additionally, it analyzes finance-specific SLMs, including FinBERT, BloombergGPT, FinGPT, and InvestLM, assessing their effectiveness in sentiment analysis, financial forecasting, and document summarization. The review highlights key research gaps, including the lack of standardized benchmarks, bias and hallucination issues, limited adaptation to real-time financial data, and insufficient multi-modal integration. Addressing these challenges is crucial for improving the reliability and applicability of SLMs in financial decision-making.**

***Index Terms*—Small Language Models (SLMs), financial data, quantization, pruning, QLoRA optimization, Retrieval-Augmented Generation (RAG), hallucinations.**

I. Introduction

Large Language Models (LLMs) have transformed Natural Language Processing (NLP) through their advanced capability to understand and reason alongside generating human-level text throughout different general-domain operations. LLMs achieve high performance levels across various applications which include code writing [3], math problem solving [4], dialogue [7], common sense reasoning [8], and symbolic reasoning [11]. The ability of LLMs to support question-answering chatbots and automation applications has turned into one of the main use cases over the past several years [1].

The main limitation of these approaches stems from their extensive parameter scale, including hundreds of billion parameters, which leads to high operational expenses when running the models at full capacity. This paper addresses these challenges by focusing on Small Language Models (SLMs), which are designed to deliver high performance with reduced computational demands. The importance of SLMs lies in their ability to operate on limited hardware, making them suitable for specialized domains like finance, where data privacy, accuracy, and efficiency are paramount. By leveraging techniques like knowledge distillation and RAG, SLMs provide a promising solution to enhance financial decision-making processes while minimizing resource consumption.

II. Preliminary Knowledge

A. Small Language Models

Small Language Models (SLMs) represent a practical solution against large language models (LLMs) because they provide effective training and inference functions that operate on limited hardware systems.e.g., weekly generation performance of 7B parameter models requires only a single consumer-grade NVIDIA RTX 4090 GPU with 24 GB of memory while utilizing 8-bit Adam training techniques [34]. Small Language Models (SLMs) feature as general-purpose language models with parameter counts below 8B. A maximum of 13B parameters in a model is regarded as exceptional when authors publish their findings [25]. The table 1 highlights how smaller, finance-specific models like FinMA-7B and fine-tuned Mistral-7B outperform larger general-purpose models like GPT-4 in financial tasks. Notably, Mistral-7B fine-tuned on financial data achieves superior results in FiQA-SA, FinArg, and SC tasks, demonstrating that domain-specific tuning enhances performance more effectively than sheer model size. [1]

TABLE I
F1-score Performance Comparison: Small vs. Large Language Models on Financial Datasets [1]

Model	Size (B)	FPB	FiQA-SA	FinArg	SC
FinMA-7B	7	**86.0**	**79.2**	27.5	45.3
GPT-4	Large	-	75.7	65.8	-
Mistral-7B	7	**86.6**	**85.5**	**85.2**	**95.5**
GPT-4	Large	-	75.7	65.8	-

B. Approaches to Create SLMs

There is a need for a series of mitigation strategies and optimization designs to enhance efficiency, decrease computational requirements, and increase reliability in developing SLMs. It begins with focus on the SLM refinement methodologies that should serve to keep performance and interpretability such that the resulting SLM will allow for effective real world application.

1) Optimization Techniques: Optimization techniques enhance SLM efficiency by reducing memory usage, computational costs, and model size while preserving accuracy. The following methods are commonly used:

- **Quantization:** Quantization techniques play a crucial role in reducing model size and processing requirements, thereby enhancing speed and efficiency in deep learning applications. By lowering the precision of model weights and activations, these methods facilitate significant reductions in memory usage and computational demands while maintaining acceptable performance levels. Weight and Activation Quantization Techniques:
  - **DilateQuant** [58]: Introduces Weight Dilation (WD) to reduce activation ranges, allowing for easier quantization without compromising accuracy.
  - **ASER** [50]: Employs Error Reconstruction and Activation Smoothing to minimize quantization errors in large language models, achieving low-bit quantization with preserved accuracy.
  - **AMED** [40]: Utilizes mixed-precision quantization, dynamically adjusting bit allocation based on hardware constraints such as memory bandwidth, processing power, and hardware-level support for lower precision formats, thereby enhancing computational efficiency without significant performance degradation.
- **Pruning:** Pruning language models is a crucial technique for enhancing their deployment in resource-constrained environments. By reducing the model size and complexity, pruning allows for efficient inference without significantly compromising performance. Various methods have been developed to achieve effective pruning, each with unique strategies and outcomes. Pruning Techniques :
  - **Activation-Based Pruning** [35]: This method identifies and removes weights with minimal contributions to neuron outputs based on activation statistics, leading to significant reductions in model size while maintaining performance.
  - **Post-Training Pruning** [26]: Techniques for pruning large language models (LLMs) without retraining have been introduced, enabling one-shot pruning to reduce resource consumption while maintaining efficiency.
  - **Layer Pruning** [20]: Pruning specific layers of LLMs can reduce memory and inference time significantly, with studies showing up to 30% layer reduction with negligible performance loss.
- **Progressive Learning:** Progressive learning is a dynamic approach that enables models to incrementally enhance their knowledge while retaining previously acquired information. This methodology is particularly beneficial in scenarios where models must adapt to new tasks without succumbing to catastrophic forgetting. Orca [28], a 13B SLM, is trained using a progressive learning approach that overcomes the limitations of imitation learning, where it learns to imitate the reasoning process of large foundation models (LFMs) such as GPT-4. Improved training signals can be observed by employing different solution strategies for various tasks, potentially distinct from those used by larger models, which can enhance smaller LMs' reasoning abilities [25].
- **Knowledge Distillation (KD) [39]:** Knowledge Distillation (KD) is a powerful technique that enables smaller models (students) to learn from larger models (teachers), enhancing their performance while maintaining efficiency. Recent advancements in KD have focused on addressing the inherent challenges posed by the capacity gap between teacher and student models, leading to innovative strategies that improve knowledge transfer.
  - **Noisy Feature Distillation** [12]: This method enhances robustness by guiding students to key pixels through spatial attention mechanisms.
  - **Speculative Knowledge Distillation** [18]: This adaptive approach addresses distribution mismatches, enhancing the quality of knowledge transfer.
  - **Chain-of-Thought Knowledge Distillation**: This method focuses on transferring both the knowledge and reasoning process (chain of thought) from a large teacher model to a smaller student model. The student is fine-tuned on the teacher's generated CoT to learn reasoning abilities for a specific task, given the smaller model's limited capacity [37].
  - **Reasoning Distillation:** Rather than mimicking the teacher's outputs, the student model is trained to replicate the intermediate reasoning steps of the teacher. This method uses a ”Decompositional distillation” strategy [38], where complex problems are broken into simpler sub-questions. A 3B model distilled using multi-step reasoning outperforms larger 11B and 6B models on the GSM8K test set [25].
  - **LoRA (Low-Rank Adaptation) [43]:**LoRA and its variant qLoRA (quantized LoRA) are innovative techniques in natural language processing (NLP) that enhance the efficiency of fine-tuning large language models (LLMs). These methods significantly reduce the number of trainable parameters, thereby lowering computational and memory requirements while maintaining performance.
    - **Parameter Efficiency** [59]: LoRA introduces low-rank adapters to linear layers, allowing for fine-tuning with fewer parameters, which is crucial for resource-limited environments
    - **qLoRA** [60]: qLoRA extends LoRA by quantizing the low-rank adapters, further reducing memory usage while preserving model performance. This approach enables the deployment of multiple specialized models on a single GPU.

2) Mitigation Strategies:

- **Retrieval-Augmented Generation (RAG):** A method that enhances the capabilities of Large

Language Models (LLMs) by integrating external knowledge, improving accuracy and reducing hallucinations. [14] This is achieved through a process that typically involves three main stages: retrieval, generation, and augmentation. [15] During retrieval, relevant information is sourced from external data stores based on a user query. [15] The generation phase then uses the retrieved information along with the original query to create a response. [15] RAG systems often use techniques such as indexing, query and embedding optimization, and vector databases to improve retrieval.The approaches for using Retrieval-Augmented Generation (RAG) in Large Language Models (LLMs) are categorized into three types: Naive RAG, Advanced RAG [17], and Modular RAG [23]. Naive and Advanced RAG approaches are widely adopted in practice due to their simplicity and low development cost. Generally, Naive RAG selects the highest cosine similarity results from a vector database and supplies the context as inputs for LLMs [21].Four common challenges of retrieval-augmented generation in LLMs: noise robustness, negative rejection, information integration, and counterfactual robustness [19].While the categorization of RAG into Naive, Advanced, and Modular approaches is accurate, the paper overlooks critical limitations of Naive RAG in the financial domain, such as its susceptibility to noisy retrievals from unstructured financial reports [19].

– **Explanation Tuning:**A fine-tuning technique aimed at improving the model's ability to generate explanations for its predictions. This process involves training the model on a dataset where the objective is not only to make accurate predictions but also to provide interpretable reasoning behind those predictions.A model is fine-tuned to provide human-understandable explanations for its predictions or decisions. [30]The goal is to improve the interpretability and trustworthiness of AI systems by aligning the model's outputs with the reasoning or justification that makes sense to humans.Orca1 [28] learns from rich explanation traces signal allowing it to overcome the limitations of instruction tuning [24].

TABLE II
PERFORMANCE COMPARISON OF SLM TECHNIQUES

Technique Covered	Key Findings
QLoRA [65]	Fine-tuning 4-bit quantization
RAG, QLoRA, Fine-tuning [66]	Improves QA-performance
QLoRA, Pruning, Quantization [67]	Speeds up inference

The table 2 summarizes the performance and efficiency trade-offs of different Small Language Model (SLM) optimization techniques. QLoRA enables memory-efficient fine-tuning, while RAG enhances retrieval-based model adaptation. Pruning and quantization further improve inference speed and reduce computational costs, making these techniques essential for deploying efficient SLMs in real-world applications. These findings highlight the impact of model compression and retrieval strategies on optimizing performance and resource utilization.

3) Hybrid Approach: The hybrid system combines effective features from the RAG model and the Fine-tuned Question Type model in a single approach (Fig. 1). The RAG model provides domain and answer context for generating results. To ensure minimal hallucination, the domain match is verified for those with a low hallucination level according to the CRAG benchmark [27] results from the movie domain. The final answer is returned when the information comes from the domain and meets the validity requirements. If the answer is incorrect, the Fine-tuned Question Type Model analyzes the question. This model specializes in rejecting invalid premise questions and also produces ”I don't know” responses for other types of questions. The system labels responses as invalid under two conditions: either when the response details an invalid question or when a JSON processing issue occurs.

C. Evaluation Metrics in LLMs

1) Text Generation and Machine Translation Metrics:

– **Perplexity (P_{er}) [48]:** Perplexity (Per) serves as a crucial metric in evaluating the performance of language models, particularly in the context of fine-tuning (FT). A lower perplexity indicates that the model generates tokens with higher confidence, reflecting its understanding of the language structure.

– **BLEU (Bilingual Evaluation Understudy) Score [49]:** The BLEU (Bilingual Evaluation Understudy) score is a widely recognized metric for assessing the quality of machine-generated translations by measuring n-gram overlap with reference texts. A higher BLEU score indicates greater similarity between the generated text and reference translations, suggesting improved accuracy in translation outputs.

– **TER (Translation Edit Rate) [49]:** The Translation Edit Rate (TER) is a crucial metric for assessing the quality of machine translation outputs, where a lower TER indicates fewer edits needed to align the generated text with a reference translation. Recent advancements have enhanced the traditional TER, making it more reflective of human judgment and applicable across various languages and contexts.

– **ROUGE (Recall-Oriented Understudy for Gisting Evaluation) Score [51]:** The ROUGE score is a critical metric for evaluating the quality of generated summaries by measuring their similarity to reference texts. A higher ROUGE score indicates greater alignment with the reference context, which is essential for assessing the effectiveness of summarization systems. The following sections elaborate on key

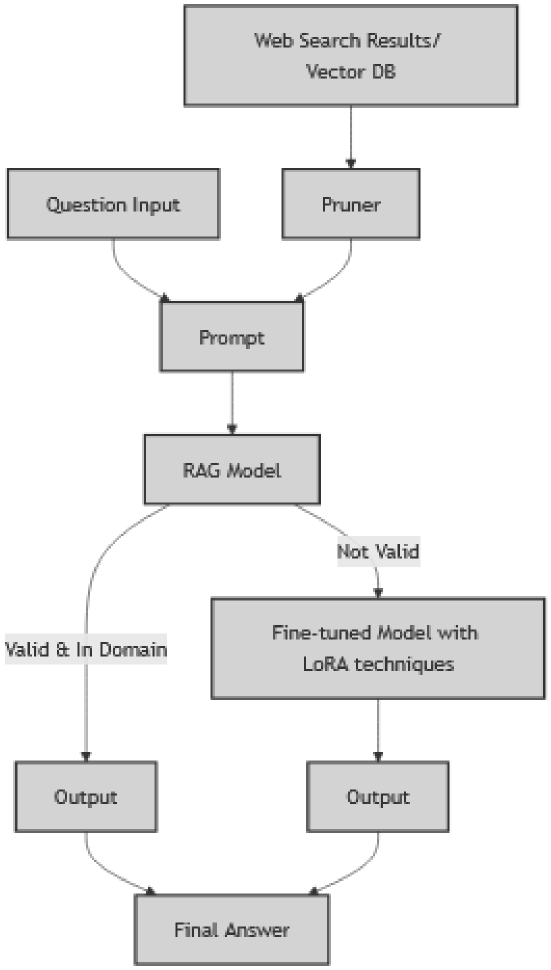

Fig. 1. Hybrid Approach for creating SLMs [21]

aspects of ROUGE scores and their implications in text summarization.

– **Hallucination Score [21]:** The Hallucination Score is a critical metric for evaluating the accuracy of generated text, particularly in the context of natural language generation (NLG) and text-to-image models. This score quantifies the degree of factual inaccuracies or fabricated content, with a lower score indicating better performance.

2) Regression Metrics (Numerical Prediction Tasks):

– **Mean Absolute Error (MAE) [56]:** The Mean Absolute Error (MAE) is a crucial metric in finance for assessing the accuracy of predictive models. It quantifies the average absolute differences between predicted and actual values, providing a clear interpretation of model performance.

– **Root Mean Squared Error (RMSE) [57]:** The Root Mean Squared Error (RMSE) is a critical metric for assessing model performance, particularly in contexts where large errors are more detrimental than smaller ones. RMSE calculates the square root of the average of squared differences between predicted and observed values, thus emphasizing larger discrepancies.

3) Information Retrieval and Search Metrics:

– **Precision [52]:** Precision, defined as the fraction of retrieved documents that are relevant, is a critical metric in evaluating information retrieval systems. It reflects the effectiveness of a system in returning relevant results while minimizing irrelevant ones. Various methodologies have been proposed to enhance precision, particularly in contexts like Technology-Assisted Review (TAR) and systematic literature reviews.

– **Recall [53]:**The concept of recall in information retrieval is crucial for evaluating the effectiveness of systems designed to retrieve relevant documents. Recall is defined as the fraction of relevant documents successfully retrieved from a larger set.

4) Prediction and Classification Metrics:

– **Accuracy [54]:** Accuracy in predictive modeling is a critical measure that reflects how often predictions align with actual outcomes. However, recent research emphasizes the importance of not only achieving high accuracy but also ensuring that the rationales behind predictions are valid.

– **F1-score [55]:** The F1-score is a crucial metric in evaluating models, particularly in scenarios with imbalanced datasets, such as fraud detection. It effectively balances precision and recall, making it particularly useful when the cost of false positives and false negatives is significant.

5) Financial Benchmark Framework: A system used to evaluate financial language models (FinLMs) by assessing their performance on tasks like sentiment analysis, entity recognition, forecasting, and risk management. It includes task-specific datasets, standardized evaluation metrics (accuracy or ROUGE), and protocols for consistent assessments. This framework helps compare models' capabilities, identify performance gaps, and guide improvements in financial applications.

– **FinBen [64]:** A proposed framework for evaluating financial language models (FinLMs) across key tasks like sentiment analysis, financial entity recognition, document summarization, risk analysis, and financial forecasting. It standardizes the evaluation of FinLMs by offering tailored datasets, metrics, and protocols to assess model performance in financial applications.

– **PIXIU [62]:** The PIXIU benchmark provides standardized evaluation metrics and instruction data for assessing the performance of financial language models (FinLMs) on a range of tasks. These tasks include sentiment analysis, financial entity recognition, financial document summarization, and market forecasting.

D. Challenges faced by Language Models

1) Hallucination and Creativity in LLMs [31]: The main disadvantage of using LLM-based solutions is the generation of hallucinatory responses which appear as incorrect or deceptive information [2]. The predictions from the later layers experience uncertainty when determining the next sequence token [9]. The research [9] [10] establishes that hallucinations occur unpredictably but share the same model parameters as creative outputs. Research presented in [9] creates a mathematical model which defines LLM hallucinations through probability theory and information theory methods while proving their characterization through low sequential token probability measures. Self-supervised learning uses metrics such as ROUGE and TER and BLEU to establish its link with hallucinations [29]. The quality evaluation of fine-tuned text involves sequential log-probabilities per token as well as the following metrics.

2) Data security: Deploying Small Language Models (SLMs) on private server enhances data security for financial data, ensuring compliance with regulatory requirements and minimizing exposure to external threats. This approach is crucial for fraud detection, risk analysis, and algorithmic trading, where data privacy is a priority. [47]

E. Domain-Specific SLMs

Task-agnostic SLMs provide a wide range of knowledge functionality, but industry vertical-based SLMs deliver optimal results in specialized fields while performing industry-specific operations [25].

1) Medical Domain: BioGPT [36] is an SLM in the medical domain that applies generative data augmentation to the PubMedQA dataset with additional fine-tuning. It achieves better results than few-shot GPT-4. Low-Rank Adaptation (LoRA) [4] enables an effective fine-tuning process, extracting essential data characteristics for BioGPT development [25].

2) Legal Domain: LawyerLLaMA [44] is one of the earliest attempts at building LLMs in the Chinese legal domain. ChatLaw is another Chinese legal domain expert model, a LoRA fine-tuned version of Ziya-LLaMA-13B trained on 937k Chinese National Law examples. ChatLaw outperforms both GPT-4 and LawyerLLaMA [33].

3) Retail Domain: SLMs demonstrate excellent capabilities when used for prompt learning in domain-focused text classification tasks across the retail industry [32]. T5-base with 220M parameters functions as an SLM evaluated through prompt-based model fine-tuning techniques. It is particularly effective in few-shot learning environments [25].

4) Finance Domain [41]: Small language models specialized for financial operations are designed to solve three primary tasks: market prediction [42], financial report analysis, and customer communication response generation. These models focus on efficiency while retaining performance. They leverage fine-tuning and transfer learning for domain-specific tasks, providing cost-effective solutions for financial applications [46].

III. Related Works

A. Language Models in Finance Domain

The financial domain is characterized by significant numerical data, data transformations, abbreviations, and domain-specific definitions. Some recent industrial products and use cases in this domain include [5]:

- Automated financial statement analysis
- Personalized narrative generation for financial reports
- Financial forecasting and prediction
- Risk management and compliance
- Audit processes

LLMs have demonstrated advanced capabilities in providing insights, identifying trends, and conducting assessments in the financial domain. Notable models such as FINBERT [13], introduced in 2022, showcase the adaptation of LLMs to financial applications. Innovations continue with BloombergGPT [6], a 50-billion-parameter model trained on extensive financial domain data, making it one of the largest and most powerful financial-specific LLMs to date. FinGPT [17] is an open-source language model designed for the finance domain, providing researchers and practitioners with accessible and transparent resources to develop Financial Language Models (FinLLMs). Its potential applications include robo-advising, algorithmic trading, and low-code development, serving as stepping stones for users. Small language models (SLMs) have been explored for task-specific training, such as FinBERT [13], but pretraining and instruction fine-tuning have primarily been investigated for larger models in the 65B range, like InvestLM and BloombergGPT [6] [9].InvestLM is trained using the CFA (Chartered Financial Analyst) exam questions and SEC (Securities and Exchange Commission) filings [9].

Table 3 presents an overview of various finance-specific Small Language Models (SLMs) and their respective capabilities. These models have been developed to address different Natural Language Processing (NLP) tasks in the financial domain. FinBERT specializes in sentiment analysis, financial entity recognition, and classification tasks, making it well-suited for financial text analysis. BloombergGPT and FLANG extend these capabilities by incorporating named entity recognition and document classification, enhancing their ability to process structured financial information. InvestLM focuses on sentiment analysis and financial text classification, improving the accuracy of financial predictions. Additionally, FinMA and FinGPT emphasize financial document summarization and question-answering, enabling more efficient extraction of insights from large financial datasets. These models collectively highlight the growing role of SLMs in automating financial text processing and decision-making.

TABLE III
CAPABILITIES OF FINANCE-SPECIFIC SMALL LANGUAGE MODELS

Finance-Specific LM	Model Capabilities
FinBERT [13]	Sentiment analysis Financial entity recognition Financial classification tasks
BloombergGPT [6]	Sentiment analysis Named entity recognition Question answering
FLANG [62]	Sentiment analysis Named entity recognition Document classification
InvestLM [9]	Sentiment analysis Financial text classification
FinMA [62]	Sentiment analysis Financial document summarization Question answering
FinGPT [5]	Financial document summarization Question answering

Table 4 provides a detailed comparison of various Small Language Models (SLMs) and their corresponding dataset sources and parameter sizes. BloombergGPT, a large-scale financial model, is trained on FinPile with 50 billion parameters, making it one of the most extensive models in this category. FinBERT and FLANG, both open-source models, are significantly smaller, with 110 million parameters each, trained on Fin.PhraseBank. InvestLM, based on the LLaMA architecture, has 658 million parameters and is trained on CFA(Chartered Financial Analyst),SEC(U.S. Securities and Exchange Commission) financial datasets. FinMA and FinGPT, also LLaMA-based models, have varying parameter sizes, with FinMA ranging from 7 billion to 13 billion parameters and FinGPT from 2 billion to 8 billion parameters, trained on PIXIU and FinQA/FinRed datasets, respectively. Other notable models include TinyLlama, Apple-OpenELM, and Microsoft-phi, which have parameter sizes ranging from 270 million to 3 billion. The Google-gemma model, part of the Gemini family, has 2 billion parameters, although its training dataset is unspecified. This table highlights the diversity in model sizes and training datasets, reflecting different trade-offs between computational efficiency and task-specific performance.

TABLE IV
MODEL PARAMETERS COUNT,TRAINED DATASET OF LANGUAGE MODELS

Language Models	Dataset	Parameter count
BloombergGPT(close)	FinPile [6]	50B [62]
FinBERT(open)	Fin.PhraseBank [13]	110M [62]
FLANG (open)	-	110M [62]
InvestLM(LLaMA-open)	CFA,SEC [9]	65B [61]
FinMA(LLaMA-open)	PIXIU [62]	7B & 13B [61]
FinGPT(open)	FinQA,FinRed [5]	7B & 13B [61]
Google-gemma(Gemini-open)	-	2B [61]
TinyLlama(LLaMA-open)	-	1.1B [61]
Apple-OpenELM	RefinedWeb,Pile [63]	270M - 3B [61]
Microsoft-phi	-	1B - 3B [61]

Table 5 shows a study conducted in [61] various Small Language Models (SLMs) on financial data set using both zero-shot and few-shot learning approaches. In the zero-shot setting, a simple instruction prompt was included, whereas the few-shot setting incorporated five in-context examples crafted by a business researcher. Among the tested models, OpenELM-270M demonstrated the best accessibility in terms of GPU efficiency, inference speed, and output readability, while Phi models required higher GPU resources. Performance analysis using ROUGE scores highlighted the superiority of few-shot models, with the highest ROUGE-1 score of 0.2683 achieved by the Gemma-2B few-shot model, while the lowest was 0.1699 for the Phi-1B zero-shot model. Similarly, the highest ROUGE-2 score was 0.0429 for the Phi-2B few-shot model, with the lowest at 0.0125 for the Phi-1B zero-shot model. The top-performing models were predominantly few-shot models, including Gemma-2B, TinyLlama-1.1B, OpenELM-1.1B, and OpenELM-270M in ROUGE-1 evaluations. Since SLMs do not match the performance of larger models, the study did not compare results to models like ChatGPT-4o, Claude, or LLaMA. It suggests that research should focus on developing higher-quality financial question-answering datasets and integrating knowledge graphs and Retrieval-Augmented Generation (RAG) pipelines to create more consumer-ready models.

TABLE V
MODEL MEMORY REQUIREMENTS AND INFERENCE TIME FOR SMALL LANGUAGE MODELS [61]

Model	GPU (GiB)	RAM (MB)	Avg.Inf Time(s)
(1)Apple-OpenELM-270M	2.2	642.2977	5.64
(2)Apple-OpenELM-450M	3.7	588.7348	7.32
(3)Apple-OpenELM-1.1B	8.2	765.3945	9.89
(4)Apple-OpenELM-3B	13.6	473.3031	14.60
(5)Microsoft-phi-1B	8.2	759.8051	7.28
(6)Microsoft-phi-1.5B	8.2	670.2625	7.30
(7)Microsoft-Phi-2B	10.3	410.8238	7.07
(8)Google-gemma-2B	9.5	792.9766	6.68
(9)TinyLlama-1.1B	8.3	721.0668	5.65

It is evident that there is a need to develop more accurate small language models, as existing models exhibit limitations in accuracy and fail to meet the varied requirements of task-specific applications in finance domain.

IV. CONCLUSION

The study highlights the viability of Small Language Models (SLMs) as an efficient alternative to Large Language Models (LLMs) for financial data applications. By leveraging advanced optimization techniques such as

TABLE VI
ROUGE EVALUATION FOR SMALL LANGUAGE MODELS ON FINANCIAL DATASET (MEAN ZERO-SHOT * MEAN FEW-SHOT) [61]

Model	ROUGE-1	ROUGE-2
(1)	0.2497 * 0.2533	0.0362 * 0.0379
(2)	0.2303 * 0.2487	0.0285 * 0.0359
(3)	0.2533 * 0.2579	0.0373 * 0.0401
(4)	0.2469 * 0.2445	0.0363 * 0.0372
(5)	0.1699 * 0.2251	0.0125 * 0.0280
(6)	0.2164 * 0.2515	0.0244 * 0.0364
(7)	0.2390 * 0.2485	0.0402 * **0.0429**
(8)	0.2013 * **0.2683**	0.0250 * 0.0428
(9)	0.1970 * 0.2626	0.0282 * 0.0390

quantization, QLoRA fine-tuning, and knowledge distillation, SLMs achieve a balance between computational efficiency and predictive accuracy. The integration of Retrieval-Augmented Generation (RAG) further enhances model reliability by reducing hallucinations and improving contextual relevance. Despite their advantages, challenges such as data security, bias mitigation, and hallucination control remain critical areas for further research. Future advancements should focus on integrating multiple data sources, improving evaluation benchmarks, and developing ethical guidelines to ensure responsible AI deployment in financial applications. This research underscores the transformative potential of SLMs in enhancing financial analysis while maintaining cost-effectiveness and operational efficiency.

V. FUTURE WORK

Future research on financial SLMs should focus on improving data security through advanced encryption methods.The detection of hallucinations in financial text should be enhanced by implementing adversarial training alongside self-supervised learning methods. Predictive capabilities will get improved through the integration of multiple data sources which include numerical data and textual financial records. Standard evaluation benchmarks for financial NLP tasks should be implemented because a performance assessment standard ensures practical business usability. User organizations need guidelines that address ethical problems and control biases to enable responsible AI application deployments in financial services.

REFERENCES

[1] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, LLMs in e-commerce: A comparative analysis of GPT and Llama models in product review evaluation," *Natural Language Processing Journal*, vol. 6, p. 100056, 2024.

[2] S. Roychowdhury, A. Alvarez, B. Moore, M. Krema, M. P. Gelpi, P. Agrawal, F. M. Rodríguez, Á. Rodríguez, J. R. Cabrejas, P. M. Serrano et al., Hallucination-minimized data-to-answer framework for financial decision-makers," in 2023 IEEE International Conference on Big Data (BigData), IEEE, 2023, pp. 4693–4702.

[3] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. Ponde, J. Kaplan, H. Edwards, Y. Burda, et al., Evaluating Large Language Models Trained on Code," *arXiv preprint arXiv:2107.03374*, 2021,p.4.

[4] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, et al., Solving Quantitative Reasoning Problems with Language Models," in Advances in Neural Information Processing Systems, vol. 35, pp. 3843–3857, 2022.

[5] X.-Y. Liu, G. Wang, H. Yang, and D. Zha, FinGPT: Democratizing internet-scale data for financial large language models," in *NeurIPS Workshop on Instruction Tuning and Instruction Following*, 2023,p.2 .

[6] 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 ,p.3.

[7] A. Glaese, N. McAleese, M. Trebacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger, M. Chadwick, P. Thacker, et al., Improving alignment of dialogue agents via targeted human judgements," *arXiv preprint arXiv:2209.14375*, 2022, p.1.

[8] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, et al., PaLM: Scaling Language Modeling with Pathways," arxiv:2204.02311, 2022, p.17.

[9] M. Lee, A mathematical investigation of hallucination and creativity in GPT models," *Mathematics*, vol. 11, no. 10, p. 2320, 2023.

[10] X. Jiang, Y. Tian, F. Hua, C. Xu, Y. Wang, and J. Guo, A survey on large language model hallucination via a creativity perspective," arXiv e-prints, p.5 arXiv–2402, 2024.

[11] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, and D. Zhou, "Self-Consistency Improves Chain of Thought Reasoning in Language Models," in The Eleventh International Conference on Learning Representations 2023.

[12] C. Shi, Y. Hao, G. Li, and S. Xu, "Knowledge Distillation via Noisy Feature Reconstruction," Expert Systems with Applications,Dec. 2024.

[13] A. H. Huang, H. Wang, and Y. Yang, "Finbert: A large language model for extracting information from financial text," Contemporary Accounting Research, vol. 40, no. 2, pp. 806–841, 2023.

[14] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University; Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University; College of Design and Innovation, Tongji University, 2023.

[15] I. Iaroshev, R. Pillai, L. Vaglietti, and T. Hanne, "Evaluating Retrieval-Augmented Generation Models for Financial Report Question and Answering," School of Business, University of Applied Sciences and Arts Northwestern Switzerland, 4600 Olten, Switzerland; Institute for Information Systems, University of Applied Sciences and Arts Northwestern Switzerland, 4600 Olten, Switzerland, 2025.

[16] X. Ma, Y. Gong, P. He, et al., "Query Rewriting in Retrieval-Augmented Large Language Models," in *EMNLP 2023*, Association for Computational Linguistics, Singapore, 2023.

[17] H. S. Zheng, S. Mishra, X. Chen, et al., "Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models," in *ICLR 2024*, 2024.

[18] W. Xu, R. J. Han, Z. Wang, L. T. Le, D. Madeka, L. Li, W. Y. Wang, R. Agarwal, C.-Y. Lee, and T. Pfister, "Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling," Oct. 2024.

[19] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," Chinese Information Processing Laboratory, State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, China; University of Chinese Academy of Sciences, Beijing, China, 2025.

[20] D. K. Thennal, T. Fischer, and C. Biemann, "Large Language Models Are Overparameterized Text Encoders," Oct. 2024.

[21] X. Chen, L. Wang, W. Wu, Q. Tang, and Y. Liu, "Honest AI: Fine-Tuning 'Small' Language Models to Say 'I Don't Know', and Reducing Hallucination in RAG," Independent Researchers, 2025.

[22] Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, and W. Chen, "Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy," *Findings of the Associ-

ation for Computational Linguistics: EMNLP 2023*, Association for Computational Linguistics, Singapore, pp. 9248–9274, 2023.

[23] W. Yu, D. Iter, S. Wang, Y. Xu, M. Ju, S. Sanyal, C. Zhu, M. Zeng, and M. Jiang, "Generate Rather than Retrieve: Large Language Models are Strong Context Generators," *The Eleventh International Conference on Learning Representations*, 2023.

[24] H. Gupta, S. A. Sawant, S. Mishra, M. Nakamura, A. Mitra, S. Mashetty, and C. Baral, "Instruction Tuned Models Are Quick Learners," 2023.

[25] S. Subramanian, V. Elango, and M. Gungor, "Small Language Models (SLMs) Can Still Pack a Punch: A Survey," Amazon, January 13, 2025.

[26] P. Zhao, F. Sun, X. Shen, P. Yu, Z. Kong, Y. Wang, and X. Lin, "Pruning Foundation Models for High Accuracy without Retraining," Oct. 2024.

[27] X. Yang, K. Sun, H. Xin, Y. Sun, N. Bhalla, X. Chen, S. Choudhary, R. D. Gui, Z. W. Jiang, Z. Jiang, L. Kong, B. Moran, J. Wang, Y. E. Xu, A. Yan, C. Yang, E. Yuan, H. Zha, N. Tang, L. Chen, N. Scheffer, Y. Liu, N. Shah, R. Wanga, A. Kumar, W. T. Yih, and X. L. Dong, "CRAG – Comprehensive RAG Benchmark," https://arxiv.org/abs/2406.04744, 2024.

[28] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, and A. Awadallah, "Orca: Progressive Learning from Complex Explanation Traces of GPT-4," 2023.

[29] G. Christopoulos, "The Impact of Language Family on D2T Generation in Under-Resourced Languages," Master's thesis, Utrecht University, 2024.

[30] C. Jeong, "Fine-Tuning and Utilization Methods of Domain-Specific LLMs," arXiv preprint arXiv:2401.02981, 2024.

[31] S. Roychowdhury, M. Krema, B. Moore, X. Lai, D. Effedua, and B. Jethwani, "FiSTECH: Financial Style Transfer to Enhance Creativity without Hallucinations in LLMs," Corporate Data and Analytics Office (CDAO), Accenture LLP, USA.

[32] H. Luo, P. Liu, and S. Esping, "Exploring Small Language Models with Prompt-Learning Paradigm for Efficient Domain-Specific Text Classification," 2023.

[33] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., "The Llama 3 Herd of Models," arXiv preprint arXiv:2407.21783, 2024.

[34] J. Zhao, Z. Zhang, B. Chen, Z. Wang, A. Anandkumar, and Y. Tian, "Galore: Memory-Efficient LLM Training by Gradient Low-Rank Projection," arXiv preprint arXiv:2403.03507, 2024.

[35] M. Hussien, M. Afifi, K. K. Nguyen, and M. Cheriet, "Small Contributions, Small Networks: Efficient Neural Network Pruning Based on Relative Importance," Oct. 2024.

[36] Z. Guo, P. Wang, Y. Wang, and S. Yu, "Dr. Llama: Improving Small Language Models on PubMedQA via Generative Data Augmentation," arXiv abs/2305.07804, 2023.

[37] L. C. Magister, J. Mallinson, J. Adamek, E. Malmi, and A. Severyn, "Teaching Small Language Models to Reason," arXiv abs/2212.08410, 2022.

[38] K. Shridhar, A. Stolfo, and M. Sachan, "Distilling Reasoning Capabilities into Smaller Language Models," 2023.

[39] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," International Journal of Computer Vision, vol. 129, no. 6, pp. 1789–1819, Mar. 2021.

[40] M. Kimhi, T. Rozen, A. Mendelson, and C. Baskin, "AMED: Automatic Mixed-Precision Quantization for Edge Devices," Mathematics,, Jun. 2024.

[41] H. P. Josyula, "Predictive Financial Insights with Generative AI: Unveiling Future Trends from Historical Data," Senior Product Manager, FinTech, Independent Researcher.

[42] Y. A. Reddy, "Predictive Modeling of Financial Market Trends Using Advanced Machine Learning Algorithms," Vignan Institute of Technology and Science.

[43] J. E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv abs/2106.09685, 2021.

[44] A. Sajith and K. C. R. Kathala, "Is Training Data Quality or Quantity More Impactful to Small Language Model Performance?," arXiv preprint arXiv:2411.15821, 2024.

[45] J. Choe, K. Noh, N. Kim, S. Ahn, and W. Jung, "Exploring the Impact of Corpus Diversity on Financial Pretrained Language Models," arXiv preprint arXiv:2310.13312, 2023.

[46] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili, et al., "A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage," Authorea Preprints, 2023.

[47] C. Zhang, J. Xia, B. Yang, H. Puyang, W. Wang, R. Chen, I. E. Akkus, P. Aditya, and F. Yan, "Citadel: Protecting Data Privacy and Model Confidentiality for Collaborative Learning," in *Proceedings of the ACM Symposium on Cloud Computing*, Seattle, WA, USA, Nov. 2021.

[48] S. Audrad, J. Sullivan, and O. Bennett, "Optimizing Pretraining Datasets for Large Language Models Through Recursive Perplexity Correlations," *arXiv preprint arXiv:2409.01678*, 2024.

[49] P. Koch, "sacRebleu: Metrics for Assessing the Quality of Generated Text," *arXiv preprint arXiv:2404.12345*, 2024.

[50] W. Zhao, Y. Shi, X. Lyu, W. Sui, L. Shen, and Y. Li, "ASER: Activation Smoothing and Error Reconstruction for Large Language Model Quantization," Nov. 2024.

[51] N. Sanchan, "Comparative Study on Automated Reference Summary Generation using BERT Models and ROUGE Score Assessment," *Journal of Current Science and Technology*, May 2024.

[52] W. Kusa, G. Peikos, M. Staudinger, et al., "Normalised Precision at Fixed Recall for Evaluating TAR," *arXiv preprint arXiv:2408.12345*, 2024.

[53] S. Singhania, S. Razniewski, and G. Weikum, "Recall Them All: Retrieval-Augmented Language Models for Long Object List Extraction from Long Documents," *arXiv preprint arXiv:2405.12345*, 2024.

[54] L. Tang, M. Ma, and X. Peng, "Beyond Accuracy: Ensuring Correct Predictions With Correct Rationales," *arXiv preprint arXiv:2410.12345*, 2024.

[55] F. Chettiar, "Integrating Autoencoders with Local Outlier Factor and Isolation Forest for Effective Fraud Detection in Imbalanced Datasets," *International Journal For Science Technology And Engineering*, 10 Oct. 2024.

[56] S. M. Robeson and C. J. Willmott, "Decomposition of the mean absolute error (MAE) into systematic and unsystematic components," *PLOS ONE*, 17 Feb. 2023.

[57] S. Reiter and S. W. R. Werner, "Interpolatory model order reduction of large-scale dynamical systems with root mean squared error measures," *arXiv preprint arXiv:2403.12345*, 13 Mar. 2024.

[58] X. Liu, Z. Li, and Q. Gu, "DilateQuant: Accurate and Efficient Diffusion Quantization via Weight Dilation," Sep. 2024.

[59] D. Cherniuk, A. Mikhalev, and I. Oseledets, "Run LoRA Run: Faster and Lighter LoRA Implementations," arXiv preprint arXiv:2312.03717, Dec. 2023. [Online]. Available: https://arxiv.org/abs/2312.03717.

[60] J. Zhao, Y. Wang, W. Abid, G. Angus, A. Garg, J. Kinnison, A. Sherstinsky, P. Molino, T. Addair, and D. Rishi, "LoRA Land: 310 Fine-tuned LLMs that Rival GPT-4, A Technical Report," Apr. 2024. [Online]. Available: https://arxiv.org/abs/2404.14597.

[61] T. R. Kosireddy, J. D. Wall, and E. Lucas, "Exploring the Readiness of Prominent Small Language Models for the Democratization of Financial Literacy," in *Michigan Technological University, 1400 Townsend Drive, Houghton, Michigan, United States of America*, pp. 2-9, Oct. 2024.

[62] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, and J. Huang, "PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance," p. 2, Oct. 2023.

[63] S. Mehta, M. H. Sekhavat, Q. Cao, M. Horton, Y. Jin, C. Sun, I. Mirzadeh, M. Najibi, D. Belenko, P. Zatloukal, and M. Rastegari, "OpenELM: An Efficient Language Model Family with Open Training and Inference Framework," *Apple*,p. 2, 2024,

[64] Q. Xie, W. Han, Z. Chen, R. Xiang, X. Zhang, Y. He, M. Xia, D. Li, Y. Dai, D. Feng, Y. Xu, H. Kang, Z. Kuang, C. Yuan, K. Yang, Z. Luo, T. Zhang, Z. Liu, G. Xiong, Z. Deng, Y. Jiang, Z. Yao, H. Li, Y. Yu, G. Huh, J. Huang, X.-Y. Liu, A. Lopez-Lira, B. Wang, Y. Lai, H. Wang, M. Peng, and S. Anania, "FinBen: A Holistic Financial Benchmark for Large Language Models,",pp. 2-5, 2024.

[65] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," May 2023.

[66] 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," Feb. 2024.

[67] "Efficiency Breakthroughs in LLMs: Combining Quantization, LoRA, and Pruning for Scaled-down Inference and Pre-training," Mar. 2024. Available: https://www.marktechpost.com/2024/03/28/efficiency-breakthroughs-in-llms-combining-quantization-lora-and-pruning-for-scaled-down-inference-and-pre-training/