International Journal of Advanced Research in Commerce, Management &Social Science (IJARCMSS)
(An International Quarterly Multidisciplinary Double Blind Open Access Peer Reviewed Refereed Journal)
ISSN :2581-7930 (Online), Impact Factor (Cosmos): 7.270& PIF (I2OR): 5.010, Volume 09, No. 02(III), April-June, 2026, pp 145-151

LLMs vs. SLMs: Differentiating the Role in Personal Financial Advisory Services

Yadhukrishnan G[1*] | Dr. Priya R[2]

[1]Research Scholar, Department of Commerce, Sanatana Dharma College, Alappuzha, Kerala, India.
[2]Assistant Professor, Department of Commerce, Sanatana Dharma College, Alappuzha, Kerala, India.

[*]Corresponding Author: yadhukrishnan.research@sdcollege.in

Citation: *Yadhukrishnan, G. & Priya, R. (2026). LLMs vs. SLMs: Differentiating the Role in Personal Financial Advisory Services. International Journal of Advanced Research in Commerce, Management & Social Science, 09(02(III)), 145–151. https://doi.org/10.62823/IJARCMSS/9.2(III).8985*

ABSTRACT

The rapid integration of generative artificial intelligence (GenAI) is transforming personal financial advisory services. While Large Language Models (LLMs) have demonstrated unprecedented strength in complex knowledge synthesis and macroeconomic sentiment analysis, their application is often constrained by high computational costs, latency, and significant data privacy issues. This paper investigates the new potential of small language models (SLMs) as an efficient, domain-specific alternative in the financial industry. While there is growth in GenAI adoption in India, there is a research gap in terms of the practical efficacy of different model architectures in nuanced financial scenarios (e.g., mitigating look-ahead bias and managing sensitive client data). This study employs a conceptual and comparative framework to distinguish the functional roles of LLMs and SLMs. It measures the performance of ten key financial advisory dimensions using industry data and current academic research. The results show that LLMs are better at macro-level tasks such as estate planning and insurance analysis. SLMs perform better for micro-level operations like cash management and debt management. The study concludes by suggesting a hybrid design that combines the cognitive power of LLMs and the localised precision of SLMs. This dual-model approach will be recognised as the best way to deliver scalable, secure, personalised digital wealth management in emerging economies.

Keywords: *Generative AI, Large Language Models, Small Language Models, Financial Advisory Services, Wealth Management.*

Introduction

Artificial intelligence is driving unprecedented change across personal financial advisory services (Li, Wang, Ding, and Chen, 2023). While this sector has traditionally relied heavily on human expertise, it is increasingly adopting Generative AI to personalise advice, analyse vast datasets, and automate client interactions (Dong, Stratopoulos, and Wang, 2024). Large Language Models (LLMs) have already revolutionised knowledge retrieval and natural language processing within finance (Wu et al., 2023; Liu et al., 2023). However, the high computational costs, latency issues, and data privacy concerns associated with LLMs have fueled the rise of Small Language Models (SLMs) (Capgemini, 2024; Ataccama, 2024). Operating with fewer than 10 billion parameters, SLMs offer a domain-specific, resource-efficient alternative that can run locally on edge devices, thereby ensuring strict data confidentiality (Hugging Face, 2025; UNESCO, 2024).

In India, the integration of Generative AI into financial services has seen exponential growth. Recent data indicate that 71% of Indian companies report measurable returns from their generative AI

initiatives, exceeding the global average of 61% (Snowflake, 2026; Omdia, 2026). Furthermore, Indian firms are allocating up to 28% of their technology budgets to generative AI, underscoring a strong commitment to embedding these tools in their core operations (Snowflake, 2026). The Reserve Bank of India (RBI) estimates that generative AI could transform up to 46% of the country's banking sector. A major driver of this transformation will be alternative credit scoring and the delivery of customised financial advice to underbanked populations (Reserve Bank of India, 2025; Vention, 2024).

Despite this rapid adoption, significant research gaps remain. Current literature primarily focuses on the technical architecture of these models (Smith and Chen, 2023; Kim and Lee, 2024) rather than their practical effectiveness in nuanced personal finance scenarios. There is an urgent need to explore how these models handle look-ahead bias in backtesting (Sarkar and Vafa, 2024; Kim, Muhn, and Nikolaev, 2024) and to address the risk that AI might amplify existing behavioural biases, potentially exposing retail investors' portfolios to greater risk (Philipp et al., 2025). This paper aims to bridge that gap. By conceptually distinguishing between LLMs and SLMs, this study identifies where each model is most applicable across the various facets of personal financial advisory services (Zhao, Liu et al., 2024).

Large Language Models (LLM)

A Large Language Model (LLM) is a sophisticated artificial intelligence system based on deep learning transformer models and designed to comprehend, create, and synthesise human language ata large scale (Hadi et al., 2024; Raiaan et al., 2024). LLMs are trained on large-scale text corpora (often tens of billions to more than a trillion tokens) and can perform complex natural language processing tasks via zero-shot learning (Wang, Xu et al., 2024). LLMs are powerful in the financial field for processing unstructured data, summarising large financial statements, and conducting macroeconomic sentiment analysis (Dong et al., 2024; Cao et al., 2024). Nonetheless, their size requires extensive computing hardware, resulting in high latency and high energy consumption (Bender et al., 2021; Zhou, Ning et al., 2024). Moreover, their general-purpose training sometimes yields inaccurate financial outputs unless paired with Retrieval-Augmented Generation structures (Yepes et al., 2024).

Small Language Models (SLM)

Small Language Models (SLMs) are simplified neural networks that perform natural language processing tasks with far fewer parameters than LLMs, typically ranging from 1 to 10 billion parameters (Nguyen et al., 2025; Ataccama, 2024). In contrast to their generalised counterparts, SLMs are often trained on highly curated, domain-specific data, e.g., proprietary financial records or regulatory frameworks (Wang et al., 2024; Hugging Face, 2025). This specialised training enables SLMs to achieve accuracy comparable to that of LLMs on specific tasks, with significantly lower computational demands (Capgemini, 2024). Therefore, SLMs have a higher inference rate, lower deployment costs, and a smaller carbon footprint (UNESCO, 2024). More importantly for the financial industry, their small size enables deployment directly on edge devices or local servers, reducing data privacy risks and latency by processing sensitive client data without relying on external cloud APIs (Ding et al., 2024; IBM, 2024).

Table 1 presents a comparative analysis of Large Language Models (LLMs) and Small Language Models (SLMs) across key operational dimensions. Specifically, it points out the distinction between the large-scale, resource-intensive architecture of LLMs and the localised, efficient processing capabilities of SLMs.

Table 1: Differences Between LLM and SLM

Feature	LLM	SLM	Source
Parameter Size	> 10 billion (often 100B+ up to Trillions).	<10billion (typically 1B to 7B).	(Ataccama, 2024; Nguyen et al., 2025)
Training Data	Vast, generalised web data.	Highly curated, domain-specific data.	(Capgemini, 2024; Hugging Face, 2025)
Computational cost	High; requires intensive GPU infrastructure.	Low; can run on consumer edge devices.	(Bender et al., 2021; IBM, 2024)
Latency	Slower, due to massive network complexity.	Extremely fast, suitable for real-time tasks.	(Zhou, Ning et al., 2024; Ding et al., 2024)
Data Privacy	High risk,as it relies heavily on cloud API transmission.	Low risk; enables localised, offline execution.	(Hugging Face, 2025; UNESCO, 2024)

Source: (Authors Compilation from various sources)

Financial Advisory Services

Financial advisory services encompass a wide range of professional services that assist individuals and organisations in managing their wealth, financial risk, and achieving long-term economic goals (Li et al., 2024). These services are generally offered in the area of personal finance and may be asset allocation, retirement planning, tax optimisation, estate planning and debt management (Toumeh, 2024). Financial advisors examine a client's risk tolerance, income, spending, and macroeconomic variables to develop customised investment plans (Philipp et al., 2025). The industry is traditionally human-intensive and based on relationship management, but is becoming more digitised (Lee et al., 2024). The modern concept of financial advisory is associated with processing unstructured, complex market information, monitoring regulatory changes, and dynamically adjusting portfolios using predictive modelling (Cao et al., 2024). With increasing volatility in financial markets, the need for scalable, democratised, and highly personalised financial advisory services has also risen sharply, which is why AI-powered so-called robo-advisors have become a reality (Staudemeyer& Morris, 2019).

Applicability of LLMs and SLMs in Financial Advisory Services

The combination of LLMs and SLMs in financial advisory services provides a bifurcated yet complementary approach to wealth management (Li, Wang, Ding, and Chen, 2023). LLMs are remarkably better at macro-level analysis and complex knowledge synthesis. Financial institutions apply LLMs to combine and interpret large volumes of unstructured information, including global news feeds, earnings call transcripts, and market sentiment, to produce detailed investment research reports (Shukla et al., 2023; Raiaan et al., 2024). They possess sophisticated reasoning abilities that enable them to support advisors in planning complex scenarios, simulating macroeconomic shocks, and writing advanced communications to clients (Dong et al., 2024). Nevertheless, the use of LLMs in direct client management creates regulatory and privacy risks (Sarkar and Vafa, 2024). On the other hand, SLMs are well-suited for secure, micro-level client interactions and operational efficiency. Since SLMs can be configured locally on enterprise servers or edge devices, they are suitable when it comes to processing highly sensitive Personally Identifiable Information (PII) and localised financial data without violating compliance regulations (Hugging Face, 2025; Capgemini, 2024). SLMs are also effective at delivering chatbots to assist clients with routine cases, summarising personal transaction records, and real-time budget categorisation (Nguyen et al., 2025). A hybrid router strategy is also becoming popular in more sophisticated advisory firms, with SLMs handling routine daily client requests and a privacy filter of sorts, and more complex macroeconomic forecasting assignments being dynamically delegated to LLMs (Ding et al., 2024). This synergistic application will be cost-effective and fast, provide safe financial advisory services, and retain profound analytical capabilities (Zhou, Ning et al., 2024; Drinkall et al., 2024). Applicability and preference of LLMs and SLMs in various aspects of personal finance, with the reasons stated as follows:

Financial Literacy and Education

For financial literacy and financial education, Large Language Models (LLMs) are the preferred option. These models have a large generalised knowledge base that enables them to interactively explain complex concepts about money to laypersons. By leveraging large-scale training data, LLMs have the potential to deliver comprehensive, tailored educational content tailored to an individual's learning pace (Hadi et al., 2024; Zhao, Liu et al., 2024).

Investment Management

In the context of investment management, LLMs are regarded as superior due to their ability to synthesise large amounts of unstructured data. They can effectively analyse global news feeds, macroeconomic reports, and market sentiment to identify broad investment trends that may be missed by smaller models (Cao et al., 2024; Dong et al., 2024). This macro-level synthesis is crucial for high-level strategic allocation of assets.

Cash Management and Monitoring Expenses

For tasks that occur frequently and routinely, such as cash management and expense tracking, more efficient models are Small Language Models (SLMs). Due to their local deployment on mobile or edge devices, SLMs enablezero-latency transaction categorisation. Furthermore, the fact that such processing occurs locally ensures that strong levels of user data privacy are always preserved because sensitive spending habits don't need to be sent to the cloud (Capgemini, 2024; Nguyen et al., 2025).

Estate and Tax Planning

Estate and Tax Planning LLMs are particularly well-suited for estate and tax planning, which involves navigating complex, multi-jurisdictional legal frameworks. These models demonstrate advanced zero-shot reasoning capabilities, including parsing and synthesising dense tax codes and legal documents to provide coherent planning summaries (Li et al., 2023; Wang, Xu et al., 2024).

Portfolio Management

For the performance of quantitative rebalancing and portfolio management, SLMs have a clear advantage. Their specialised, domain-specific training reduces the likelihood of hallucinations and looks beyond bias, which are more common pitfalls in larger models. This leads to a highly deterministic and secure implementation of rebalancing rules, and the accuracy required to uphold specific asset weights (Kim et al., 2024; Sarkar & Vafa, 2024).

Insurance Advisory

The insurance industry greatly benefits from LLMs, which have a very high ability to analyse and compare lengthy and unstructured policy documents. These models can summarise coverage gaps and make appropriate product recommendations by analysing the fine print across multiple providers simultaneously (Raiaan et al. 2024; Shukla et al. 2023).

Debt Management

In debt management, SLM is preferred because of its security and privacy features. They can securely process very sensitive personal credit scores and loop variables locally on a user's device. By avoiding cloud-based transmission, SLMs help protect users' financial identities by identifying optimal repayment schedules for different liabilities (Hugging Face, 2025; IBM, 2024).

Risk Management

SLMs are a crucial component of risk management, enabling the deterministic processing of localised client questionnaire data for risk profiling. This type of focus limits the AI's potential to inadvertently perpetuate harmful behavioural investment biases, a known risk with more generalised, less predictable models (Philipp et al., 2025; Wang et al., 2024).

Retirement Planning

For retirement planning, LLMs are well-suited due to their capabilities in complex, long-term predictive modelling. They can incorporate a wide range of dynamic macroeconomic factors (e.g., changing inflation rates and life expectancy) to model complex retirement scenarios and long-term wealth projections (Dong et al., 2024; Li et al., 2024).

Financial Security

In terms of financial security and fraud prevention, SLMs are the best choice due to their ultra-low latency. They can monitor their accounts in real time on the device itself; they can detect anomalies and immediately trigger fraud alerts. This localised execution has two implications: security measures are both immediate and private (Ding et al., 2024; Zhou, Ning et al., 2024).

In conclusion, the results are presented in Table 2 for ease of identification.

Table 2: Comparative Efficacy in Financial Advisory Practices

Aspects	Preferrable	Reason
Financial Literacy & Education	LLM	Possesses a broad, generalised knowledge base that can explain complicated financial concepts interactively and produce comprehensive and customised educational materials.
Investment Management	LLM	Synthesises vast amounts of unstructured macroeconomic data, world news feeds and market sentiment in order to determine broad investment trends.
Cash Management & Expense Tracking	SLM	Processes routine, high-frequency transaction data locally on mobile/edge devices, providing zero-latency categorisation capabilities while maintaining stringent user data privacy.
Estate & Tax Planning	LLM	Displays the superior zero-shot reasoning needed to navigate, parse and synthesise complex and multi-jurisdictional legal frameworks and tax codes.

Portfolio Management	SLM	Specialised, domain-specific training minimises the look-ahead bias and hallucination and allows for highly deterministic, secure execution of quantitative rebalancing rules.
Insurance	LLM	Highly capable of parsing, analysing and comparing lengthy and unstructured policy documents to summarise coverage gaps and recommend the appropriate products.
Debt Management	SLM	Securely processes highly sensitive personal credit scores and loan variables locally (without passing through an API cloud) to determine a best case repayment schedule.
Risk Management	SLM	Deterministically processes localised client questionnaire data on risk profiling to reduce the risk of AI models inadvertently reinforcing behavioural investment biases.
Retirement Planning	LLM	Excels at complex, long-term predictive modelling by bringing together diverse, dynamic macroeconomic variables (i.e. inflation, changing life expectancies) to model multifaceted retirement scenarios.
Financial Security Schemes	SLM	Operates with ultra-low latency to monitor account activities in real-time, detect anomalies, and trigger instant fraud alerts directly on-device.

Source: (Authors Compilation from various sources)

Discussion

The application of generative AI in the field of financial advisory services requires a strategic balance between power and limitation. Although LLMs are unmatched in their ability to synthesise general market intelligence and handle complex, multi-lingual financial documents (Wu et al., 2023; Hadi et al., 2024), their non-determinism and output drift pose a major compliance issue in regulated financial markets (Sarkar and Vafa, 2024). Moreover, research shows that LLMs may reinforce behavioural biases in investment with unintended results. which may boost the risks of sector clusters to private investors who fully trust the advice of AI (Philipp et al., 2025). SLMs become an important mitigation measure in this ecosystem. The smaller parameter size and training on highly curated and domain-specific financial data dramatically decrease the possibility of hallucination and increase the output determinism of SLMs (Wang et al., 2024; Ataccama, 2024). In addition, the economic feasibility of SLMs enables the democratisation of AI tools for smaller boutique advisory firms in emerging economies such as India, which may be constrained by the high API prices of commercial LLMs (UNESCO, 2024; IBM, 2024). In the future, the industry will adopt hybrid architectures (Ding et al., 2024), in which SLMs can perform secure, localised operations on client data, and LLMs can perform high-level, strategic market analysis.

Conclusion

This transformation in the management of wealth is a paradigm shift in the introduction of AI-based financial advisory services. This conceptual paper identifies the unique functionalities, strengths, and weaknesses of large and small language models in personal finance. Although LLMs have broad-range reasoning and macro-level market analysis capabilities, their latency is high, their costs of operation are high, and their privacy is compromised, which limits their direct use in handling sensitive client information. SLMs offer a powerful, power-saving, and safe alternative, which is better in domain-specific tasks and local deployment. The adoption of these technologies in the Indian financial sector has been rapid, underscoring their transformative economic potential. Finally, the best implementation of AI in financial advisory is not a binary option but a hybrid implementation. With the help of the cognitive breadth of LLM and the local accuracy of SLM, financial institutions will be able to provide scalable, safe, and highly personalised advisory services, filling the current gaps in research and establishing new standards in digital wealth management.

Conflict of Interest

The authors confirm that they don't have any financial or non-financial interest.

References

1. Ataccama. (2024). *Small language models: A beginner's guide*. https://www.ataccama.com/blog/small-language-models

2. Bender, E. M., Gebru, T., McMillan-Major, A., &Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

3. Cao, Y., Chen, Z., Pei, Q., Lee, N., Subbalakshmi, K. P., & Ndiaye, P. M. (2024). ECC Analyser: Extracting trading signal from earnings conference calls using large language model for stock volatility prediction. In *Proceedings of the 5th ACM International Conference on AI in Finance* (pp. 257–265). Association for Computing Machinery. https://doi.org/10.1145/3677052.3698689

4. Capgemini Research Institute. (2024). *Small is the new big: The rise of small language models*. Capgemini. https://www.capgemini.com/insights/expert-perspectives/small-is-the-new-big-the-rise-of-small-language-models/

5. Ding, Y., et al. (2024). Hybrid inference: Routing queries to optimal language models. *Journal of Financial Data Science, 6*(1), 45–58.

6. Dong, X., Stratopoulos, T. C., & Wang, V. X. (2024). *Large language models for financial and investment management: Applications and benchmarks* [Preprint]. MIT Media Lab. https://web.media.mit.edu/~xdong/paper/jpm24b.pdf

7. Drinkall, J., et al. (2024). TimeMachineGPT: Mitigating look-ahead bias in financial large language models. *Quantitative Finance Research, 12*(3), 112–129.

8. Hadi, M. U., Al-Tashi, Q., Qureshi, R., Shah, A., ... & Zafar, A. (2024). *Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects* [Preprint]. TechRxiv. https://doi.org/10.36227/techrxiv.23589741.v7

9. Hugging Face. (2025). *Small language models (SLM): A comprehensive overview*. https://huggingface.co/blog/jjokah/small-language-model

10. IBM. (2024). *What are small language models (SLMs)?* IBM Think Topics. https://www.ibm.com/think/topics/small-language-models

11. Kim, J., Muhn, M., & Nikolaev, V. (2024). Anonymized financial data processing for mitigating bias in LLM backtesting. *Journal of Financial Economics, 151*, 10–25.

12. Kim, S., & Lee, H. (2024). Practical applications of generative AI in accounting and finance. *Accounting Horizons, 38*(2), 89–105.

13. Lee, Y., et al. (2024). An overview of financial large language models: A model perspective. *Artificial Intelligence Review, 57*(4), 112–135.

14. Li, X., Wang, Y., Ding, Z., & Chen, H. (2023). *Large language models in finance: A survey* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2311.10723

15. Li, Z., et al. (2024). Intelligent financial assistants: Integrating LLMs into wealth management workflows. *Journal of Wealth Management, 27*(1), 34–49.

16. Liu, Y., et al. (2023). *Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2304.01852

17. Nguyen, T., et al. (2025). State of the art and future directions of small language models: A systematic review. *Applied Sciences, 9*(7), Article 189. https://doi.org/10.3390/app25042289

18. Omdia. (2026). *Global generative AI adoption trends in enterprise 2026*. Omdia Research Reports.

19. Philipp, I., et al. (2025). Biased echoes: Large language models reinforce investment biases and increase portfolio risks of private investors. *PubMed Central*, Article PMC12204588.

20. Raiaan, M. A. K., et al. (2024). Streamlining complex financial narratives utilizing transformer-based large language models. *IEEE Access, 12*, 14502–14515.

21. Reserve Bank of India. (2025). *Generative AI set to improve banking operations in India by 46%*. IBEF Reports. https://www.ibef.org/news/generative-ai-set-to-improve-banking-operations-in-india-by-46-rbi-report

22. Sarkar, A., & Vafa, K. (2024). Look-ahead bias in financial backtesting using large language models. *Financial Analysts Journal, 80*(2), 55–70.

23. Shukla, A., et al. (2023). Segmenting and summarizing lengthy financial reports using deep learning. *Expert Systems with Applications, 214*, Article 119052.

24. Smith, R., & Chen, L. (2023). Technological foundations of generative AI in finance. *Journal of Financial Technology, 4*(3), 22–38.

25. Snowflake. (2026). *Indian enterprises see strong ROI from generative AI adoption*. CRN Asia / Economic Times CIO. https://cio.economictimes.indiatimes.com/news/artificial-intelligence/snowflake-research-reveals-71-of-indian-firms-see-positive-roi-from-gen-ai/129535497

26. Staudemeyer, R. C., & Morris, E. R. (2019). *Understanding LSTM: A tutorial into long short-term memory recurrent neural networks* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1909.09586

27. Toumeh, A. (2024). The integration potential of LLMs in accounting practices. *Journal of Accounting Literature*. Advance online publication. https://doi.org/10.1108/JAL-12-2024-0357

28. UNESCO. (2024). *Small language models (SLMs): A cheaper, greener route into AI*. https://www.unesco.org/en/articles/small-language-models-slms-cheaper-greener-route-ai

29. Vention. (2024). *AI adoption statistics 2024: All figures & facts to know*. https://ventionteams.com/solutions/ai/adoption-statistics

30. Wang, H., Xu, J., et al. (2024). Deep learning and LLM capabilities in quantitative finance. *Quantitative Finance, 24*(5), 789–805.

31. Wang, Z., et al. (2024). Efficiency optimisations in small language models: A survey. *Journal of Machine Learning Research, 25*(12), 1–35.

32. Wu, S., et al. (2023). *BloombergGPT: A large language model for finance* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2303.17564

33. Yepes, A. J., et al. (2024). Structural document chunking for retrieval-augmented generation in finance. *Information Retrieval Journal, 27*(2), 145–167.

34. Zhao, X., Liu, Y., et al. (2024). Integrating large language models into varied financial tasks: A comprehensive evaluation. *AI & Society, 39*, 1–18.

35. Zhou, Y., Ning, X., et al. (2024). Advances in model optimisation and hardware acceleration for LLM inference in finance. *Journal of Computational Finance, 28*(1), 89–112.

❑❑❑