2025; 1-1 | pp.1-10

International Journal of Economic Dynamics and Finance
Volume 1, Issue 1, July-September 2025
(An Academians Publishers of Research & Academic Resources)

ISSN: XXXX-XXXX

Original Research Article

Received: 18-07-2025 **Accepted**: 20-08-2025 **Published**: 24-09-2025

Behavioral Integration in Personal Finance Large Language Models: A Data-Centric Framework for Efficient and Trustworthy Financial Advisory Systems

A.R. Mostafa*[1]

Abstract: The rapid advancement of large language models (LLMs) is transforming the landscape of automated financial advisory, offering the potential for scalable, personalized, and data-driven guidance. However, current solutions are constrained by significant computational demands, insufficient adaptation to individual user behavior, and the persistence of systematic biases. In this paper, we propose a comprehensive data-centric framework that fundamentally rethinks the training and deployment of personal finance LLMs. Central to our approach is the integration of behavioral finance principles at every stage of the model development pipeline, including an innovative four-phase chain-of-thought generation process that explicitly incorporates users' psychological states and emotional cues into financial reasoning. We constructed an extensive dataset of 19,000 real-world personal finance queries across eight diverse domains—ranging from debt management to retirement planning—and used it to fine-tune a Qwen-3-8B model. Our multi-faceted evaluation, combining held-out testing, blind LLM-jury studies, and cost-benefit analysis, demonstrates that the resulting 8B-parameter model matches or surpasses the performance of models 2–4 times larger (14–32B parameters) in factual accuracy, response fluency, and degree of personalization, while reducing operational costs by 80%. This framework directly tackles the pitfalls of current agentic and general-purpose approaches, notably their architectural complexity, high maintenance burden, and poor real-world efficiency. By showing that principled behavioral integration and robust data design can substitute for brute-force computational scaling, our work paves the way for practical, accessible, and genuinely trustworthy financial AI systems applicable to a broad range of users and scenarios.

Keywords: *Large Language Models, Personal Finance, Behavioral Finance, Chain-of-Thought Reasoning, Financial Advisory, Data-Centric*

[1]*Independent Scholar*

Introduction

The intersection of artificial intelligence and financial services has witnessed unprecedented growth, with automated advisory systems becoming increasingly prevalent across investment management, personal budgeting, and financial planning (Sanz-Cruzado et al., 2024; Takayanagi et al., 2023). However, the deployment of large language models in high-stakes financial domains presents unique challenges that distinguish it from general-purpose applications. Personal financial advice requires not only factual accuracy and domain expertise but also sophisticated understanding of individual psychological profiles, risk tolerance, and behavioral biases that significantly influence financial decision-making (Baker et al., 2017; Zhou et al., 2025).

Available Online:

***Corresponding Author: A.R. Mostafa**

Recent advances in LLM-based financial systems have predominantly followed two approaches: scaling existing general-purpose models to handle financial tasks through in-context learning, or developing complex agentic architectures with multiple specialized components (Okpala et al., 2025; Takayanagi et al., 2025a). While these approaches demonstrate promising capabilities on standardized benchmarks, their practical deployment faces significant constraints. Large-scale models require substantial computational infrastructure, making them prohibitively expensive for widespread deployment. Agentic systems, despite their theoretical sophistication, encounter rapid performance degradation in real-world conditions and achieve less than 25% of their anticipated returns due to integration complexity, maintenance overhead, and adaptation challenges (Meimandi et al., 2025).

A fundamental limitation of current approaches lies in their treatment of behavioral and psychological factors as secondary considerations rather than integral components of financial reasoning. Traditional financial advisory inherently involves understanding and addressing cognitive biases, emotional states, and individual risk profiles. However, existing LLM-based systems often amplify rather than mitigate these biases, leading to systematically biased advice that can increase portfolio risk and reinforce poor financial decisions (Winder et al., 2024; Zhi et al., 2025).

This paper addresses these challenges through a data-centric approach that integrates behavioral finance principles directly into the training process. Rather than relying on complex architectures or massive model scaling, we develop a principled framework for generating behaviorally-grounded supervision data that enables smaller, more efficient models to provide personalized, bias-aware financial guidance.

Research Contributions

Our work makes several significant contributions to the intersection of AI and financial services:

1. **Behaviorally-Grounded Framework:** We introduce a novel four-phase chain-of-thought generation pipeline that treats psychological cue identification as a foundational component of financial reasoning, ensuring that behavioral considerations are intrinsic to the model's decision-making process.
2. **Data-Centric Efficiency:** Through careful curation and behavioral integration, we demonstrate that an 8B parameter model can achieve performance comparable to systems 2-4 times larger while reducing operational costs by 80%.
3. **Comprehensive Evaluation:** We conduct rigorous evaluation through held-out testing, blind LLM-jury studies, and cost-benefit analysis, providing evidence for both the effectiveness and practical viability of our approach.
4. **Societal and Ethical Considerations:** We discuss the societal impact and ethical implications of deploying automated financial advisory systems, including transparency, fairness, and the need for robust mechanisms to avoid exacerbating inequalities. This is accompanied by a commitment to reproducibility and open science, with publicly available datasets, models, and evaluation protocols to facilitate future research.

Literature Review

Evolution of AI in Financial Services

The application of artificial intelligence to financial advisory has evolved through several distinct phases. Early systems employed rule-based expert systems and collaborative filtering techniques for specific domains such as loan approval and insurance

Available Online:

recommendations (Zibriczky, 2016). The advent of machine learning introduced more sophisticated approaches, including neural networks for market prediction and portfolio optimization, though these remained limited to narrow, well-defined tasks.

The emergence of large language models has fundamentally transformed the landscape, enabling natural language interaction and facilitating reasoning about complex financial scenarios. Recent studies have demonstrated LLMs' capabilities across various financial tasks, including text summarization, sentiment analysis, market forecasting, and investment recommendation (Gueta et al., 2025; Liu et al., 2025). However, comprehensive evaluations reveal significant limitations in current general-purpose models when applied to financial domains, with no single model excelling across all task categories (Matlin et al., 2025).

Challenges in LLM-Based Financial Advisory

Computational and Deployment Constraints

The pursuit of improved performance through model scaling has created significant practical barriers to deployment. State-of-the-art models require substantial computational resources, with inference costs often exceeding practical thresholds for widespread adoption (Cemri et al., 2025). This has led to a proliferation of agentic approaches that attempt to compensate for individual model limitations through complex multi-agent architectures.

Behavioral Bias Amplification

A critical concern in LLM-based financial systems is the tendency to amplify rather than mitigate human cognitive biases. Zhou et al. (2025) conducted comprehensive analysis demonstrating that LLMs exhibit significant financial biases, including anchoring, overconfidence, and representativeness heuristics. More alarmingly, fine-tuning on financial data can sometimes exacerbate these irrational tendencies, creating systems that provide systematically biased advice.

Empirical studies have documented specific manifestations of these problems, including product bias in investment recommendations (Zhi et al., 2025) and systematic increases in portfolio risk through reinforcement of geographical concentration and trend-chasing behaviors (Winder et al., 2024). These findings highlight the inadequacy of treating behavioral considerations as an afterthought in system design.

Performance-Deployment Gap

Recent analysis has revealed a significant discrepancy between benchmark performance and real-world effectiveness in financial AI systems. Meimandi et al. (2025) found that technical and cost-related factors prevent agentic financial advisors from realizing even 25% of their anticipated returns. This performance degradation is attributed to the inherent volatility of real-world conditions, integration challenges with existing systems, and the limited context that can be effectively supplied to complex multi-agent architectures.

Behavioral Finance and AI Integration

The field of behavioral finance has established robust theoretical frameworks for understanding and addressing cognitive biases in financial decision-making (Baker et al., 2017; Agrawal, 2012). These insights have found limited integration into AI systems, despite their critical importance for personalized financial advice. Traditional approaches have treated behavioral considerations as post-hoc corrections rather than fundamental components of the reasoning process.

Recent work has begun exploring more sophisticated integration approaches. Takayanagi et al. (2025a) demonstrated that users' trust and engagement are heavily influenced by the perceived persona of the advisor, rather than the accuracy of the advice. This finding suggests that effective financial AI systems must incorporate empathetic reasoning and personalized communication strategies as core capabilities rather than superficial additions.

Available Online:

Methodology

Framework Overview

Our approach centers on a data-centric framework that generates behaviorally-grounded supervision data through a structured chain-of-thought process. Rather than relying on complex architectural modifications or massive parameter scaling, we focus on creating high-quality training data that integrates financial domain knowledge with behavioral finance principles.

Behaviorally-Grounded Reasoning Chain Generation Pipeline for Personal Finance LLMs

The framework operates through four interconnected phases, each designed to capture different aspects of personalized financial reasoning. This modular approach enables independent validation and optimization of each component while maintaining coherent integration in the final reasoning chain.

Data Collection and Processing

Source Selection and Ethical Considerations

We collected authentic financial scenarios from Reddit's personal finance communities, particularly r/personalfinance, which receives millions of user queries annually. This platform offers access to complex, multifaceted scenarios that capture the intricate nature of real-world financial decision-making. To ensure ethical compliance, we exclusively utilized publicly available archived data from posts prior to June 2023, implementing aggressive de-identification procedures that removed all personally identifiable information while preserving the essential financial context.

Quality Filtering and Categorization

Our data processing pipeline employed a two-stage filtering approach: topical validity filtering to retain posts containing explicit, answerable personal finance questions, and contextual clustering to group semantically similar posts and remove near-duplicates. This process yielded 405,000 unique questions, from which we sampled 19,000 representative queries spanning eight thematic categories as detailed in our comprehensive dataset analysis.

Table 1: Dataset Distribution Across Personal Finance Categories

Category	Count	Avg Query Tokens	Avg CoT Tokens	Avg Response Tokens
Debt Management & Credit	5175	215.76	628.3	393.69
Retirement Planning	3286	198.1	648.28	407.02
Tax Planning & Optimization	3019	182.96	630.2	397.81
Investing & Wealth Building	2994	200.16	653.54	402.98
Budgeting & Cash-Flow Management	2503	221.53	628.71	394.47
Insurance & Risk Management	1035	213.86	621.53	389.65
Savings & Emergency Funds	638	177.18	652.25	382.95
Estate Planning & Legacy	196	216.9	653.47	409.06

Available Online:

The dataset spans diverse financial domains, with debt management and credit optimization representing the largest category (5,175 samples), followed by retirement planning (3,286 samples) and tax optimization strategies (3,019 samples). Each category demonstrates consistent token distributions for queries, chain-of-thought reasoning, and final responses, indicating balanced representation across financial domains.

Chain-of-Thought Generation Pipeline

Phase 1: Query Analysis and Decomposition

The initial phase addresses the inherent inconsistency and complexity of natural language financial inquiries. Many user queries contain significant redundancy, embedded emotional context, or critical information presented in unclear ways. Our query analysis component systematically decomposes user questions into essential semantic elements, identifying the primary financial conflict, key stakeholders, and critical facts necessary for resolution.

This decomposition serves multiple purposes: it eliminates conversational distractions that could bias downstream reasoning, ensures all relevant parties and constraints are considered, and provides a structured foundation for subsequent analysis phases. The modular design allows each component to focus on specific aspects while maintaining coherent integration.

Phase 2: Dual-Corpus Context Analysis

Context analysis employs a modular retrieval-augmented generation (RAG) framework built on two carefully curated corpora. The financial corpus comprises approximately 600,000 tokens from authoritative sources, including Investopedia and the Bogleheads community, covering core concepts such as retirement accounts, debt repayment strategies, and consumer policy updates. The behavioral corpus comprises 300,000 tokens of research and practitioner insights, spanning the psychology of risk, investor behavior, behavioral portfolio theory, and generational differences in financial decision-making.

Our retrieval process combines semantic search with cross-encoder reranking to identify the most relevant information chunks, followed by LLM-based synthesis to create a streamlined, source-attributed context. This approach ensures that both factual financial knowledge and behavioral insights are available for downstream reasoning while maintaining tractable context lengths.

Phase 3: Psychological Cue Identification

A distinguishing feature of our framework is the explicit identification and modeling of users' psychological states as a foundational component of financial reasoning. This phase extracts overall sentiment, primary emotions identifiable from word choice, and certainty levels present in user queries. These cues inform the selection of appropriate communication strategies and bias mitigation techniques.

The psychological analysis operates independently from financial reasoning to prevent emotional states from biasing objective analysis while ensuring that final responses can synthesize both rational and empathetic components. This separation enables the system to provide financially sound advice while acknowledging and addressing the emotional dimensions of financial decisions.

Phase 4: Response Formulation and Generation

The final phases consolidate information from all preceding analyses into comprehensive response generation instructions. This includes synthesis of financial recommendations, behavioral bias mitigation strategies, and empathetic communication approaches tailored to the user's psychological profile.

Response generation adheres to strict guidelines to ensure natural, user-friendly output that conceals the complex underlying reasoning architecture while delivering

Available Online:

actionable, personalized advice. The generated responses maintain professional standards while incorporating appropriate emotional intelligence and awareness of bias.

Phase	Purpose	Key Output
Query Analysis	Decompose user query into essential components	Primary conflict, stakeholders, financial facts
Context Analysis (Financial)	Retrieve relevant financial knowledge from 600k token corpus	Actionable financial rules and procedures
Context Analysis (Behavioral)	Extract behavioral finance insights from 300k token corpus	Cognitive biases, debiasing tactics, user-state cues
Psychological Cue Identification	Identify sentiment, emotions, and communicative intent	Overall sentiment, emotions, certainty level
Response Formulation	Synthesize instructions from all preceding phases	Consolidated directives for response generation
Response Generation	Generate final user-facing response with appropriate tone	Personalized financial advice with empathetic framing

Validation and Quality Assurance

Each phase of the generation process employs LLM-based jury evaluation to ensure quality and consistency. Multiple judge models, including Gemini-2.0-flash and GPT-4-mini, evaluate generated content using three-shot frameworks and rank multiple generation candidates. This multi-judge approach, which incorporates family diversity, helps mitigate individual model biases and ensures robust quality control throughout the pipeline.

Experimental Design

Model Training and Configuration

We fine-tuned the Qwen-3-8B model using AdamW optimization with bfloat16 precision over four epochs with a batch size of 256. Training employed a cosine learning rate schedule with maximum learning rate of 5×10^-5 and 10% linear warmup. The training split consisted of 15,600 samples, with 2,600 validation samples, requiring approximately 3 hours on a single A100 GPU.

Evaluation Methodology

Quantitative Assessment

A quantitative evaluation was conducted using a held-out dataset of 500 queries across personal finance categories. We computed BERTScore using Qwen-3-8B embeddings for semantic accuracy assessment and BLEURT scores for fluency evaluation. Ground truth responses were generated using our framework (not the fine-tuned model) and validated by independent jurors.

Qualitative Analysis through LLM-Jury Studies

To assess generalization capabilities and avoid limitations of reference-based metrics, we conducted a blind LLM-jury evaluation on 504 entirely held-out queries from subsequent time periods to prevent data contamination. We employed diverse judges from unrelated model families to minimize familial bias, using DeepSeek-v3 and Kimi-k2 with multiple independent rankings per query.

Evaluation criteria were decomposed into three orthogonal dimensions: accuracy (financial correctness), plausibility (reasoning quality), and relevance (task alignment). This decomposition avoids single scalar metrics that can reward fluent but unsafe answers while enabling targeted error analysis.

Results and Analysis

Performance Comparison

Table 2: Comprehensive Model Performance Comparison

Available Online:

Model	BERT Score	BLEURT
Gemma3-27B-IT	0.7142	0.4374
Gemma3-12B-IT	0.7139	0.439
Mistral-24B-2501	0.7133	0.4464
QWQ-32B (reasoning)	0.7069	0.4452
DeepSeek-Qwen-14B (reasoning)	0.7069	0.4513
Ours (8B)	0.7	0.46
Llama-3 8B	0.6881	0.4547
Mistral-7B v0.3	0.665	0.4501

Cost-Benefit Analysis our 8B model achieves semantic accuracy comparable to leading baselines, including Gemma3-27B and Mistral-24B, while surpassing larger models by 3-5% in human-likeness and fluency metrics. This indicates reduced deviation from ground-truth data and enhanced fluency signals compared to models twice its size.

LLM-Jury Evaluation Results

Comprehensive Evaluation Results: Performance, Cost, and Efficiency Analysis

The jury evaluation demonstrates that our behaviorally tuned 8B model achieves competitive performance across all evaluation dimensions, while significantly outperforming models of similar size. The radar visualization shows our model approaching the performance of 27-32B parameter leaders while maintaining substantial efficiency advantages.

Rank correlation analysis reveals substantial agreement between judge sets (Kendall's $\tau \approx 0.62$-0.69, Spearman's $\rho \approx 0.76$-0.83), indicating consistent evaluation quality. Relevance demonstrates the strongest alignment ($\tau = 0.691$, $\rho = 0.826$), suggesting reliable assessment of task completion and user constraint satisfaction.

Model	Size (GB)	Endpoint Cost ($/h)	GPU	Inference Time (s/query)	Total Cost ($)
QWQ-32B	65	3.8	4xL4	167.86	22.33
Gemma3-27B	46.4	2.5	1xA100	64.34	5.63
Gemma3-12B	20	1.8	1xL40S	58.26	3.67
Ours (8B)	16.4	0.8	1xL4	34.15	0.96
Mistral-24B-2501	46.1	3.8	1xA100	37.99	5.05
DeepSeek-Qwen-14B	29.5	1.8	1xL40S	54.18	3.41
Llama3-8B	16.1	0.8	1xL4	33.58	0.94
Mistral-7B	14.5	0.8	1xL4	29.15	0.82

Cost and Deployment Analysis: Demonstrating 80% Cost Reduction with Competitive Performance

Our cost analysis reveals dramatic efficiency improvements through the data-centric approach. The 8B model requires only $0.96

Available Online:

to process 504 queries compared to $22.33 for QWQ-32B, representing an 80% cost reduction while maintaining competitive performance. Average inference time of 34.15 seconds per query enables responsive financial advisory services without prohibitive infrastructure requirements.

The parameter efficiency analysis demonstrates that our model achieves the highest Borda points per billion parameters across all evaluation criteria, validating that careful supervision can substitute for scale in domain-specific applications.

Error Analysis and Limitations

Qualitative analysis reveals consistent strengths in structural organization, empathetic framing, and the extraction of user-specific details. The model reliably acknowledges emotional context before providing practical guidance, enhancing perceived helpfulness and user engagement.

Primary weaknesses include factual inaccuracies, particularly in relation to jurisdiction-specific regulations and tax details. These errors are most frequent in regulation-heavy domains (taxes, insurance) and least common in general planning tasks (budgeting, debt management). This pattern suggests that targeted retrieval for regulatory information and calculation verification would yield the highest marginal improvements.

Discussion

Implications for Financial AI Development

Our results demonstrate that behavioral integration and principled data construction can achieve performance competitive with much larger systems while dramatically reducing operational costs. This finding has significant implications for the development of practical financial AI systems, suggesting that domain-specific optimization may be more effective than general-purpose scaling for specialized applications.

The explicit modeling of psychological states as a foundational component of reasoning represents a paradigm shift from treating behavioral considerations as post-hoc corrections. This approach ensures that empathy and bias awareness are intrinsic to the model's decision-making process, leading to more trustworthy and effective financial guidance.

Advantages Over Agentic Approaches

Our data-centric framework addresses several critical limitations of complex agentic systems. By incorporating behavioral and financial knowledge directly into model weights, we avoid the complexity of integration, maintenance overhead, and real-time knowledge curation requirements that limit the effectiveness of agentic systems. This approach enables standalone deployment without complex orchestration while maintaining high-quality advisory capabilities.

Scalability and Extension

The modular framework design supports incremental extension to additional domains, jurisdictions, and behavioral dimensions. The separation of behavioral and financial analysis phases enables targeted improvements without requiring a complete system redesign. Future extensions could incorporate mixture-of-experts architectures for regional specialization or multi-modal capabilities for document analysis.

Future Directions

Global Scaling and Localization

Future research should explore optimal strategies for global deployment, including systematic market porting of US-optimized pipelines and the development of mixture-of-experts frameworks with regional specialization. Detection of regional signals such as currency symbols, policy terminology, and regulatory frameworks could enable lightweight expert modules while maintaining a shared backbone for universal financial logic.

Advanced Behavioral Integration

Available Online:

The psychological analysis component represents an initial step toward comprehensive behavioral integration. Future work could incorporate more sophisticated psychological indicators such as risk tolerance assessments, financial stress measurements, and personality-based communication preferences derived from specialized surveys or clinical datasets.

Compliance and Safety Enhancement

Rather than relying on additional supervised fine-tuning, future development should treat financial advice generation as an alignment problem, employing preference-based optimization techniques such as Direct Preference Optimization (DPO) to refine the outputs. Rule-based compliance layers can enforce regulatory fidelity, mitigate bias, and maintain tone consistency while preserving model flexibility.

Conclusion

This research establishes a data-centric framework that enables an 8B parameter model to achieve semantic fidelity and human likeness comparable to models 2-4 times larger, while reducing operational costs by 80%. These improvements stem from three synergistic components: explicit psychological cue identification, retrieval-augmented behavioral grounding, and principled data construction for supervision.

Our approach addresses critical gaps in current financial AI systems by treating behavioral considerations as foundational rather than auxiliary components of financial reasoning. The modular design supports incremental extension while maintaining cost-effectiveness and deployment viability. We acknowledge the importance of user privacy, explainability, and fairness as central to the responsible deployment of financial AI. While limitations remain in geographic scope, behavioral depth, and regulatory coverage, this work provides a robust foundation for developing practical, trustworthy financial advisory systems.

The results validate that careful behavioral integration and data-centric optimization can substitute for computational scale in domain-specific applications, offering a compelling alternative to resource-intensive scaling approaches. As financial AI systems continue to evolve, frameworks that strike a balance between efficiency, effectiveness, and ethical considerations will become increasingly important for widespread adoption and societal benefit.

Reference

Agrawal, K. (2012). A conceptual framework of behavioral biases in finance. *IUP Journal of Behavioral Finance*, 9(1), 7-18.

Baker, H. K., Filbeck, G., & Ricciardi, V. (2017). How behavioural biases affect finance professionals. *The European Financial Review*, 2017, 25-29.

Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., ... & Stoica, I. (2025). Why do multi-agent LLM systems fail? *Preprint arXiv:2503.13657*.

Gueta, A., Feder, A., Gekhman, Z., Goldstein, A., & Reichart, R. (2025). Can LLMs learn macroeconomic narratives from social media? *Preprint arXiv:2406.12109*.

Liu, Z., Guo, X., Lou, F., Zeng, L., Niu, J., Wang, Z., ... & Zhang, L. (2025). Fin-R1: A large language model for financial reasoning through reinforcement learning. *Preprint arXiv:2503.16252*.

Matlin, G., Okamoto, M., Pardawala, H., Yang, Y., & Chava, S. (2025). Finance language model evaluation (FLAME). *Preprint arXiv:2506.15846*.

Meimandi, K. J., Aránguiz-Días, G., Kim, G. R., Saadeddin, L., & Kochenderfer, M. J. (2025). The measurement imbalance in agentic AI evaluation undermines industry productivity claims. *Preprint arXiv:2506.02064*.

Okpala, I., Golgoon, A., & Kannan, A. R. (2025). Agentic AI systems applied to tasks in financial services: Modeling and model risk

management crews. *Preprint arXiv:2502.05439*.

Sanz-Cruzado, J., Richards, E., & McCreadie, R. (2024). FAR-AI: A modular platform for investment recommendation in the financial domain. In *Advances in Information Retrieval* (pp. 267-271). Springer.

Takayanagi, T., Izumi, K., Sanz-Cruzado, J., McCreadie, R., & Ounis, I. (2025a). Are generative AI agents effective personalized financial advisors? *Preprint arXiv:2504.05862*.

Winder, P., Hildebrand, C., & Hartmann, J. (2024). Biased echoes: Generative AI models reinforce investment biases and increase portfolio risks of private investors. *Social Science Research Network*.

Zhi, Y., Zhang, X., Wang, L., Jiang, S., Ma, S., Guan, X., & Shen, C. (2025). Exposing product bias in LLM investment recommendation. *Preprint arXiv:2503.08750*.

Zhou, Y., Ni, Y., Xi, Z., Yin, Z., He, Y., Yunhui, G., ... & Chai, H. (2025). Are LLMs rational investors? A study on the financial bias in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 24139-24173).

Zibriczky, D. (2016). Recommender systems meet finance: A literature review. In *International Workshop on Personalization & Recommender Systems in Financial Services*.