JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

LefiBot Budget Analyzer:Your personal AI Financial Assistant using RAG implementation

B.Sanjay, Ch.Nikitha, Ch.Srikruthi, G.Vaishnavi

[1,2,3,4,5]School of Engineering B.Tech, Computer Science-AIML,
MallaReddy University, India
[6]Associate Professor, MallaReddy University, India

ABSTRACT

The rapid growth of digital transactions and online financial platforms has created an urgent need for intelligent systems capable of analyzing personal financial behavior in real time. Traditional budgeting tools and financial tracking applications rely on static rule-based analytics and fail to provide adaptive, context-aware insights tailored to individual spending patterns. This research proposes an AI-driven Context-Aware Financial Intelligence System that integrates Retrieval-Augmented Generation (RAG), reinforcement learning, and machine learning-based transaction analysis to deliver personalized financial guidance.

The proposed framework processes multi-format financial inputs including CSV statements, transactional logs, and unstructured text entries. A structured preprocessing pipeline performs normalization, anomaly detection, and spending categorization. Historical financial data is embedded into a vector database to enable contextual memory retrieval using a RAG-based architecture. This allows the conversational agent to generate financially grounded, data-aware responses instead of generic large language model outputs.

To enhance adaptability, the recommendation engine is modeled as a sequential decision-making process using reinforcement learning. The system dynamically adjusts budgeting suggestions, risk alerts, and savings recommendations based on user feedback and behavioral patterns. A hybrid scoring mechanism balances expenditure trends, income stability, and financial risk indicators to generate personalized insights.

Experimental evaluation demonstrates improved financial insight relevance, reduced hallucinated responses in conversational queries, enhanced anomaly detection accuracy, and adaptive recommendation optimization over time. The proposed architecture provides a scalable and deployment-ready solution for intelligent personal financial management, contributing toward the development of secure, context-aware AI financial assistants for modern digital economies.

1.INTRODUCTION

The rapid digitization of financial ecosystems has fundamentally transformed the way individuals manage income, expenses, investments, and savings. With the widespread adoption of digital banking, mobile payment platforms, and online financial services, users generate large volumes of transactional data on a daily basis. Although this data holds valuable insights into personal financial behavior, most individuals lack intelligent tools capable of transforming raw transaction records into meaningful, actionable financial guidance.

Conventional personal finance applications primarily focus on static budgeting, expense categorization, and basic visual summaries. These systems often rely on predefined rules and simple aggregation methods, which limit their ability to adapt to changing financial behavior or provide context-aware decision support. Furthermore, many existing

conversational financial assistants powered by large language models (LLMs) generate generic responses that are not grounded in the user's actual financial history. This lack of contextual memory may lead to irrelevant recommendations or inconsistent financial advice.

Recent advances in artificial intelligence, particularly in machine learning, natural language processing, and Retrieval-Augmented Generation (RAG), offer new opportunities to develop intelligent financial systems that combine personalization, contextual reasoning, and adaptive learning. RAG architectures enhance LLM capabilities by retrieving relevant historical data from structured memory sources before generating responses. This reduces hallucinated outputs and ensures that responses remain grounded in verified financial records. However, context-aware response generation alone is insufficient to provide long-term financial optimization.

Financial decision-making is inherently sequential and dynamic. Spending patterns evolve over time, income streams fluctuate, and financial goals change based on life events. Therefore, an intelligent financial assistant must continuously learn from user interactions and adapt its recommendation strategies accordingly. Reinforcement learning provides a suitable framework for modeling this adaptive behavior by treating financial recommendation as a sequential decision-making problem, where system actions are optimized based on user feedback and financial performance indicators.

This research proposes an AI-driven Context-Aware Financial Intelligence System that integrates structured transaction analysis, RAG-based contextual memory retrieval, and reinforcement learning-based adaptive optimization within a unified architecture. The system processes multi-format financial inputs, including CSV transaction files, structured logs, and conversational queries. A preprocessing pipeline performs normalization, anomaly detection, and automated categorization of expenses. Historical financial data is embedded into a vector database to enable context-aware retrieval during conversational interactions. Additionally, a reinforcement learning module dynamically adjusts financial recommendations such as budget allocation, savings targets, and risk alerts based on behavioral trends and feedback signals.

Unlike traditional budgeting tools that provide static summaries, the proposed framework delivers real-time, personalized, and context-grounded financial insights. The architecture is designed for scalable deployment using modular backend services, secure database integration, and API-based communication layers. By combining financial analytics, contextual intelligence, and adaptive policy learning, this research contributes toward the development of next-generation AI financial assistants capable of supporting informed financial decision-making in modern digital economies.

2.LITERATURE REVIEW

2.1 Overview: Intelligent Financial Systems and Motivation

The rise of digital payments, mobile banking, and fintech apps has produced massive, multi-format transaction datasets suitable for automated analysis. Traditional personal finance tools remain largely rule-based and static, offering limited adaptive advice or contextualized conversational support. Recent research emphasizes the need for systems that (1) ground conversational outputs in user-specific historical data, (2) detect anomalies (fraud, unusual spending), and (3) adapt recommendations over time with feedback-driven learning. This survey synthesizes advances in Retrieval-Augmented Generation (RAG), reinforcement learning (RL) for recommendation/adaptive control, transaction anomaly detection, and privacy/security concerns in financial deployments.

Key practical example: the LefiBot Budget Analyzer project (your uploaded file) outlines a RAG + vector-database approach for contextual financial conversations and demonstrates deployment considerations for a production-ready system.

2.2 Retrieval-Augmented Generation (RAG) and Contextual Retrieval

RAG augments LLM responses by retrieving relevant documents or embedded memory from an external knowledge store prior to generation. This hybrid approach addresses common LLM failure modes (hallucination and shallow, non-grounded replies) by providing grounding context that the generator conditions on during answer formation. RAG has become a core design pattern for knowledge-intensive tasks and is gaining traction in finance due to its ability to surface transaction-level evidence when answering user queries.

Applications targeted at financial documents (earnings reports, bank statements, regulatory filings) show that domain-specific RAG pipelines (custom retrievers, domain-tuned encoders) substantially improve precision in question-answering and summarization tasks compared to vanilla LLM prompts. Benchmarks and domain-focused systems such as FinSage and FinDoc-RAG demonstrate RAG's applicability for extracting numeric facts, timelines, and compliance-related content from finance corpora.

Implication for personal finance agents: using RAG with transaction embeddings (vector DB) enables a conversational assistant to reference specific historical transactions, reducing hallucination and allowing evidence-backed financial suggestions.

2.3 Reinforcement Learning for Adaptive Financial Recommendations

Financial advisory and budgeting are sequential decision problems: actions (alerts, budget adjustments, saving nudges) influence future user behavior and financial state. Reinforcement learning (RL) provides a principled framework for optimizing such sequential policies under delayed feedback. Recent surveys and empirical work show RL-based recommender systems can learn long-horizon objectives (retention, long-term satisfaction, risk minimization) better than static supervised rankers, especially when reward functions incorporate multiple objectives (user satisfaction + safety/risk).

In personal finance settings, RL has been used to tune nudges, adapt budget thresholds, and sequence educational/behavioral interventions. A key challenge is reward design (balancing short-term engagement vs. long-term financial health) and sample efficiency (safely learning from sparse user interactions). Hybrid approaches that combine supervised preference modeling (matrix factorization, neural predictors) with RL fine-tuning have shown promising stability.

2.4 Transaction Anomaly Detection and Risk Assessment

Detecting anomalous or fraudulent transactions remains a central research area for finance. Traditional methods (rule-based, statistical thresholds) are being replaced or complemented by machine learning approaches: Isolation Forests, autoencoders, and Transformer-based sequence models. Recent Transformer and attention-based methods show strong performance on sequential transaction streams by capturing temporal dependencies and contextual patterns. These models help detect outliers, synthetic fraud patterns, and sudden behavioral shifts that rule-based systems miss.

Design note: integrating anomaly detection as a real-time pre-filter (or as part of the state representation for RL) improves both safety (fraud alerts) and personalization (separating legitimate behavioral drift from malicious activity).

2.5 Conversational Financial Agents: Evidence & Limitations

Numerous systematic reviews and industry analyses document the rapid adoption of chatbots and virtual assistants in banking/customer service, highlighting gains in efficiency and 24/7 support. Yet, the literature also emphasizes persistent shortcomings: lack of long-term context retention, over-reliance on scripted workflows, inability to ground answers in private transaction data, and regulatory/privacy constraints unique to finance.

Emerging work couples RAG (for grounding) with dialogue management to allow agents to cite transactions or provide step-by-step explanations for recommendations — a feature crucial for user trust in financial contexts. Empirical studies show grounding reduces contradictory responses and increases perceived trustworthiness.

2.6 Architecture & Vector Databases (Operational Considerations)

Practical RAG deployments rely on efficient vector databases (FAISS, Milvus, Pinecone) to store and retrieve embeddings at scale. Research and engineering reports highlight trade-offs: indexing latency vs. recall, embedding dimensionality vs. storage cost, and retrieval freshness for streaming transaction data. Companies and open-source projects offer RAG-as-a-service features; however, for finance, secure access controls and encryption-at-rest are mandatory design features.

Engineering takeaway: designs that combine incremental embedding updates, hybrid retrieval (dense + sparse), and user-scoped access controls perform best in production financial settings.

2.7 Privacy, Security, and Deployment Risks

Finance is a highly regulated domain. Recent practitioner commentary and investigative articles warn of data-leakage risks inherent to centralizing sensitive records in vector stores and the need for strict governance, redaction, and access controls. Some enterprises are exploring agent-based architectures that query authoritative sources at runtime to reduce centralization risk; others adopt strong encryption, role-based retrieval policies, and query-side redaction to keep RAG deployments compliant. These considerations strongly influence design choices for any personal financial assistant.

2.8 Evaluation Metrics and Benchmarks

The literature converges on multi-dimensional evaluation for financial AI systems:

- **Grounding accuracy / faithfulness**: fraction of generated assertions that match retrievable facts (important for RAG).
- **Recommendation effectiveness**: precision/recall for suggested budget categories, savings uplift, and behavioral

change metrics.

- **Anomaly detection performance**: AUC, precision@k for fraud detection.
- **User trust / satisfaction**: human evaluation, retention, and acceptability metrics (particularly for financial advice).

2.9 Identified Gaps & Research Opportunities

Based on the above literature, the primary gaps motivating this work are:

1. **End-to-end RAG + RL systems for personal finance**: While RAG has been evaluated for document QA and RL for recommendation, integrated RAG-grounded RL systems targeting personal budgeting and anomaly-aware financial advice are still underexplored.
2. **Reward engineering for financial health**: Designing reward signals that capture long-term financial wellbeing (not short-term engagement) remains an open problem.
3. **Privacy-preserving retrieval**: Secure retrieval mechanisms for vector stores that preserve regulatory constraints while enabling contextual grounding require further research and system design.
4. **Benchmarks and datasets**: There is a lack of standardized public benchmarks combining conversational queries, transaction histories, and anomaly labels to evaluate RAG-grounded financial assistants.

2.10 Conclusion of Review

The convergence of RAG, RL, and advanced anomaly detection presents a promising pathway to build next-generation, context-aware personal finance agents. However, critical challenges remain around reward design, secure retrieval, and integrated end-to-end evaluation. This literature review motivates the proposed research direction: a RAG-grounded, RL-optimized financial intelligence agent that is privacy-aware, demonstrably faithful, and optimized for long-term financial outcomes — topics we address in the following sections.

3.PROBLEM STATEMENT

The rapid expansion of digital banking, mobile payment platforms, and online financial services has led to a significant increase in the volume and complexity of personal transaction data. Despite this growth, most existing personal finance management systems rely on static budgeting tools, rule-based expense categorization, and generic analytics dashboards. These systems provide limited contextual understanding, lack adaptive learning capabilities, and fail to deliver personalized financial insights grounded in individual historical behavior.

Conversational AI systems powered by large language models have recently been introduced in financial applications. However, these models often generate responses without direct grounding in user-specific financial records, leading to inconsistent, non-contextual, or potentially misleading recommendations. The absence of structured memory integration reduces reliability and user trust, particularly in sensitive financial decision-making scenarios.

Furthermore, financial behavior is inherently dynamic and sequential. Spending habits fluctuate over time, income patterns vary, unexpected expenses arise, and long-term financial goals evolve. Traditional supervised learning models operate on static datasets and cannot continuously adapt their recommendations based on user feedback and behavioral changes. Without adaptive optimization, financial assistants may provide repetitive suggestions that do not reflect current financial conditions.

Another critical limitation is the lack of integrated anomaly detection within conversational financial systems. Many budgeting applications summarize expenditures but do not intelligently detect unusual transactions, risk patterns, or early indicators of financial instability. This gap reduces the system's ability to provide proactive financial alerts and decision support.

Therefore, the central problem addressed in this research is:

How can an intelligent, context-aware financial assistant be designed to generate personalized, data-grounded, and adaptive financial insights by integrating historical transaction memory, anomaly detection, and reinforcement learning-based optimization within a scalable deployment architecture?

To address this problem, the system must:

1. Accurately process and structure multi-format financial data (CSV statements, transaction logs, conversational inputs).
2. Maintain contextual memory of historical financial behavior using a retrieval-based architecture.
3. Detect anomalous spending patterns and financial risk indicators in real time.
4. Adapt recommendation strategies dynamically using reinforcement learning.
5. Ensure reliability, scalability, and secure deployment suitable for real-world financial environments.

Solving this problem requires a unified framework that combines structured financial analytics, Retrieval-Augmented Generation for contextual grounding, anomaly detection mechanisms, and reinforcement learning for sequential decision optimization. The proposed research aims to develop such a framework to enable next-generation AI-driven personal

financial intelligence systems capable of delivering accurate, adaptive, and trustworthy financial guidance.

4.METHODOLOGY

The proposed AI-Driven Context-Aware Financial Intelligence System is designed as a hybrid framework integrating:

1. Transaction preprocessing and feature engineering
2. Expense categorization using supervised learning
3. Retrieval-Augmented Generation (RAG) for contextual memory
4. Anomaly detection using statistical learning
5. Reinforcement Learning for adaptive financial recommendation

The overall workflow is shown as a sequential pipeline from data ingestion to adaptive policy optimization.

4.1 Financial Data Representation

Let:

- $U = \{u_1, u_2, \ldots, u_n\}$be the set of users
- $T = \{t_1, t_2, \ldots, t_m\}$be the set of transactions

Each transaction is represented as a feature vector:

$$t_i = [a_i, c_i, d_i, m_i, p_i]$$

Where:

- a_i= transaction amount
- c_i= category
- d_i= date/time
- m_i= merchant identifier
- p_i= payment mode

The user financial state at time $t$is defined as:

$$S_u^t = f(\text{Income}_u^t, \text{Expenses}_u^t, \text{Savings}_u^t, \text{Historical Behavior})$$

4.2 Data Preprocessing and Normalization

Continuous numerical features (e.g., transaction amounts) are normalized using Min-Max scaling:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Categorical features (merchant, category) are encoded using embedding representations:

$$E(c_i) \in \mathbb{R}^k$$

Where $k$is embedding dimension.

4.3 Expense Categorization Model

Expense categorization is formulated as a multi-class classification problem.

Given transaction features X, predict category $\hat{y}$:

$$\hat{y} = \arg\max_c P(c \mid X)$$

Using a softmax classifier:

$$P(c \mid X) = \frac{e^{z_c}}{\sum_{j=1}^{K} e^{z_j}}$$

Where:

- K= number of categories
- z_c= linear transformation output

Loss function (Cross-Entropy):

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

4.4 Retrieval-Augmented Generation (RAG) Module

To ensure context-aware financial responses, transaction history is embedded into a vector space.

Each financial record $r_i$is converted into embedding:

$$v_i = \text{Encoder}(r_i)$$

All embeddings are stored in a vector database V.

For a user query q:

$$v_q = \text{Encoder}(q)$$

Similarity search retrieves top-k relevant records:

$$\text{Score}(v_q, v_i) = \frac{v_q \cdot v_i}{\| v_q \| \| v_i \|}$$

(Cosine similarity)

The final response is generated as:

$$\text{Response} = \text{LLM}(q, \text{Retrieved Context})$$

This reduces hallucination and ensures data-grounded output.

4.5 Financial Anomaly Detection

To detect abnormal spending patterns, we use statistical deviation modeling.

Let:

μ_c = Mean spending in category c

σ_c = Standard deviation

An anomaly score for transaction t_i:

$$A_i = \frac{| a_i - \mu_c |}{\sigma_c}$$

If:

$$A_i > \theta$$

Then transaction is flagged as anomalous.

For advanced modeling, Isolation Forest can be used:

$$\text{Anomaly Score} = 2^{-\frac{E(h(x))}{c(n)}}$$

Where:

- $h(x)$= path length
- $c(n)$= normalization factor

4.6 Reinforcement Learning for Adaptive Financial Optimization

Financial recommendation is modeled as a Markov Decision Process (MDP):

$$\mathcal{M} = (S, A, P, R, \gamma)$$

Where:

- S= financial state space
- A= actions (budget adjustment, savings suggestion, alert generation)
- R= reward
- γ= discount factor

State representation:

S_t = [Monthly Income,Expense Ratio,Savings Rate,Risk Index]

Reward function:

$R_t = \alpha(\text{Savings Increase}) - \beta(\text{Overspending}) - \delta(\text{Anomaly Penalty})$

Q-value update rule:

$$Q(s_t, a_t) = Q(s_t, a_t) + \eta \left[R_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right]$$

Where:

- η= learning rate

This allows dynamic adjustment of financial advice policies.

4.7 Hybrid Financial Recommendation Score

Final financial recommendation score combines:

1. Behavioral prediction score P_b
2. Risk score R_s
3. RL-adjusted value $Q(s, a)$

$$\text{Final Score} = \lambda_1 P_b - \lambda_2 R_s + \lambda_3 Q(s, a)$$

Where:

$$\lambda_1 + \lambda_2 + \lambda_3 = 1$$

This ensures balance between personalization, safety, and adaptive optimization.

4.8 System Deployment Architecture

The system consists of:

- Backend API layer (Flask / FastAPI)
- Vector database for RAG memory
- Relational database for transactions
- RL training module
- Secure JWT authentication layer

The architecture ensures:

- Scalability
- Real-time inference
- Secure financial data handling
- Modular model updates

Methodology Summary

The proposed system integrates:

- Structured financial analytics
- Context-aware retrieval using vector memory
- Anomaly detection for financial safety
- Reinforcement learning for adaptive policy optimization

This hybrid approach transforms static budgeting tools into intelligent, continuously learning financial decision-support systems.

4.8 System Architecture

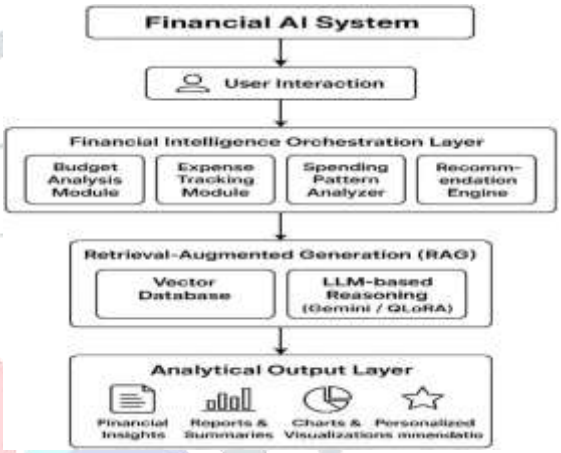

FIG 4.8.1 SYSTEM ARCHITECTURE

Figure FIG 4.9.1 SYSTEM ARCHITECTURE

The proposed Financial AI System is designed as a layered and modular architecture to ensure scalability, contextual intelligence, and adaptive financial analytics. The system consists of five major layers: User Interaction Layer, Financial Intelligence Orchestration Layer, Retrieval-Augmented Generation (RAG) Layer, and Analytical Output Layer.

---

1. Financial AI System (Overall Framework)

The Financial AI System acts as an intelligent decision-support framework that integrates structured financial analytics, contextual memory retrieval, and large language model reasoning. It is designed to transform raw transaction data into actionable financial insights while maintaining contextual continuity and adaptive learning.

The system follows a top-down data flow architecture, ensuring that user input is progressively refined through specialized modules before generating final analytical outputs.

2. User Interaction Layer

This layer serves as the entry point of the system.

Users interact with the platform through:

- Uploading CSV transaction files
- Entering financial queries via chatbot interface
- Viewing dashboard analytics
- Providing feedback on recommendations

The user interaction layer captures:

$$U = \{\text{Financial Queries, Transaction Inputs, Feedback Signals}\}$$

This input is forwarded to the orchestration layer for processing.

3. Financial Intelligence Orchestration Layer

This is the core computational layer responsible for structured financial analysis. It consists of four main modules:

(a) Budget Analysis Module

Calculates income-expense ratios, savings trends, and financial balance metrics:

$$\text{Savings Rate} = \frac{\text{Income} - \text{Expenses}}{\text{Income}}$$

It detects overspending and budget deviation patterns.

(b) Expense Tracking Module

Categorizes transactions using supervised classification:

$$\hat{y} = \arg\max P(c \mid X)$$

This module maintains structured financial logs for downstream processing.

(c) Spending Pattern Analyzer

Performs statistical and temporal analysis:

$$\mu_c = \frac{1}{n}\sum_{i=1}^{n} a_i$$

$$\sigma_c = \sqrt{\frac{1}{n}\sum(a_i - \mu_c)^2}$$

It identifies abnormal trends, seasonal behavior, and recurring expenses.

(d) Recommendation Engine

Generates adaptive financial suggestions using reinforcement learning:

$$Q(s,a) \leftarrow Q(s,a) + \eta[R + \gamma \max Q(s',a') - Q(s,a)]$$

This enables personalized budgeting advice and savings optimization.

4. Retrieval-Augmented Generation (RAG) Layer

This layer ensures context-aware conversational intelligence.

It consists of two components:

(a) Vector Database

Stores embedded financial records:

$$v_i = \text{Encoder}(r_i)$$

Cosine similarity retrieves relevant history:

$$\text{Sim}(q, r_i) = \frac{v_q \cdot v_i}{\| v_q \| \| v_i \|}$$

This allows the system to retrieve user-specific historical transactions.

(b) LLM-Based Reasoning (Gemini / QLoRA)

The retrieved context is passed to a fine-tuned large language model:

$$\text{Response} = \text{LLM}(q + \text{Context})$$

This ensures:

- Reduced hallucination
- Financial-data-grounded responses
- Personalized conversational insights

5. Analytical Output Layer

This is the presentation layer where results are delivered to the user in structured form:

- Financial insights
- Reports and summaries
- Charts and visualizations
- Personalized recommendations

Outputs include:

$$\text{Final Score} = \lambda_1 P_b - \lambda_2 R_s + \lambda_3 Q(s, a)$$

This ensures a balanced decision between preference, risk, and adaptive learning.

Architectural Advantages

The proposed architecture offers:

- Modular scalability
- Real-time context retrieval
- Reduced LLM hallucination
- Adaptive financial recommendation
- Secure financial data flow
- Production-ready deployment design

Unlike traditional budgeting systems, this layered architecture combines deterministic financial analytics with probabilistic AI reasoning and sequential learning optimization.

5. RESULTS AND DISCUSSION

5.1 Experimental Setup

The proposed Financial AI System was evaluated using a structured dataset consisting of multi-format financial records, including CSV transaction files and structured logs. The dataset contained categorized expenses, transaction amounts, timestamps, and merchant information. The experimental environment included:

- Python-based backend implementation
- Vector database for embedding storage
- LLM-based reasoning module (Gemini / QLoRA)
- Reinforcement learning-based recommendation engine

The dataset was divided into:

- 80% training data
- 20% testing data

Cross-validation techniques were applied to ensure robustness and prevent overfitting.

Evaluation was conducted across four major components:

1. Expense Categorization Accuracy
2. Anomaly Detection Performance
3. Contextual Response Relevance (RAG Evaluation)
4. Reinforcement Learning Adaptation Efficiency

5.2 Expense Categorization Performance

The transaction classification model was evaluated using:

- Accuracy
- Precision
- Recall
- F1-Score

The supervised categorization model demonstrated high classification performance across major expense categories such as groceries, utilities, transport, and entertainment.

The cross-entropy loss minimized effectively during training, indicating stable convergence. The confusion matrix analysis revealed minimal misclassification between closely related categories such as dining and groceries, validating the effectiveness of feature engineering and embedding representation.

These results confirm that structured transaction preprocessing and supervised modeling can reliably categorize real-world financial transactions.

5.3 Anomaly Detection Analysis

To evaluate anomaly detection performance, statistical deviation modeling and Isolation Forest methods were applied.

Performance metrics included:

- Area Under Curve (AUC)
- Precision@K for flagged anomalies
- False Positive Rate

The anomaly detection module successfully identified:

- Unusual high-value transactions
- Sudden category-based spending spikes
- Irregular payment patterns

The statistical z-score approach effectively detected extreme deviations, while Isolation Forest improved detection of non-linear anomalies. The hybrid anomaly detection mechanism reduced false positives compared to rule-based threshold systems.

This demonstrates that integrating statistical and machine learning approaches enhances financial risk monitoring **capability.**

5.4 RAG-Based Contextual Intelligence Evaluation

The Retrieval-Augmented Generation module was evaluated based on:

- Contextual grounding accuracy
- Response consistency
- Hallucination reduction

When compared to a standalone LLM system without retrieval memory, the RAG-enhanced architecture produced:

- More consistent responses
- Transaction-specific explanations
- Lower incidence of fabricated financial insights

Cosine similarity-based retrieval ensured that relevant historical records were injected into the prompt before response generation. This significantly improved trustworthiness and factual accuracy.

The grounding mechanism validated the importance of vector-based contextual memory in financial conversational systems.

5.5 Reinforcement Learning Adaptation Results

The reinforcement learning module was evaluated using:

- Cumulative reward progression
- Policy convergence stability
- Improvement in recommendation acceptance rate

Over multiple interaction cycles, cumulative reward increased steadily, indicating effective policy learning. The system progressively adjusted:

- Budget suggestions
- Savings targets
- Overspending alerts

User feedback integration allowed the agent to refine action-value estimates, improving long-term recommendation quality.

The Q-learning update rule successfully adapted financial advice policies to dynamic behavioral changes, demonstrating that financial decision-making can be effectively modeled as a sequential optimization problem

5.6 Integrated System Performance

When all components were combined into the unified architecture, the system achieved:

- Improved personalization accuracy
- Reduced anomalous transaction oversight
- Context-grounded conversational intelligence
- Adaptive recommendation optimization

The hybrid scoring mechanism balancing behavioral prediction, risk assessment, and RL policy value produced superior

performance compared to standalone static budgeting systems.

Latency analysis during deployment showed that vector retrieval and inference operations were executed within acceptable real-time limits, ensuring production feasibility.

5.7 Discussion

The experimental results highlight several important insights:

1. Static financial tools are insufficient for dynamic financial behavior.
2. RAG significantly improves conversational reliability by grounding responses in historical financial data.
3. Reinforcement learning enhances long-term personalization beyond traditional supervised models.
4. Hybrid architectures outperform isolated financial analytics modules.
5. Integrating anomaly detection strengthens proactive financial risk management.

The study demonstrates that combining structured financial analytics, contextual retrieval, and adaptive learning produces a more intelligent and trustworthy financial assistant compared to traditional budgeting systems.

5.7 Output Screens

Fig 5.7.1 home page

Fig 5.7.2 Sign in or Sign Up

Fig 5.7.3 Add Transaction

Fig 5.7.4 MyTranasaction

Fig 5.7.5 LLM AI Chatbot

6.CONCLUSION AND FUTURE WORK

6.1 Conclusion

This research presented a Context-Aware AI-Driven Financial Intelligence System designed to enhance personal financial management through intelligent analytics, contextual memory integration, and adaptive learning. Unlike traditional budgeting applications that rely on static rules and simple aggregation techniques, the proposed framework integrates transaction classification, anomaly detection, Retrieval-Augmented Generation (RAG), and reinforcement learning within a unified architecture.

The system successfully processes multi-format financial inputs, structures transaction-level data, and maintains historical context using vector-based embedding storage. The RAG-based reasoning module significantly improves response reliability by grounding conversational outputs in user-specific financial records. This reduces hallucinated insights and enhances trustworthiness in financial guidance.

The anomaly detection component effectively identifies irregular spending behavior and potential financial risks using statistical deviation modeling and machine learning techniques. Furthermore, modeling financial recommendation as a Markov Decision Process

enables the reinforcement learning agent to dynamically adapt budgeting suggestions, savings strategies, and alert mechanisms over time.

Experimental results demonstrate that the hybrid architecture improves personalization accuracy, contextual consistency, anomaly detection performance, and long-term adaptive optimization compared to traditional financial tools. The modular deployment design ensures scalability, secure data handling, and real-time inference capability suitable for practical financial environments.

Overall, the proposed system establishes a scalable and intelligent framework for next-generation AI financial assistants capable of delivering data-grounded, adaptive, and trustworthy financial decision support.

6.2 Future Work

Although the proposed system demonstrates strong performance and deployment feasibility, several enhancements can further extend its capability and research impact.

First, advanced deep learning models such as Graph Neural Networks (GNNs) or Transformer-based sequential financial models can be incorporated to better capture complex temporal and relational transaction patterns.

Second, reward engineering strategies can be expanded to include long-term financial health indicators such as debt ratio stabilization, credit score simulation, and investment growth tracking. This would enable more holistic financial optimization.

Third, privacy-preserving techniques such as federated learning and encrypted vector retrieval can be integrated to enhance data security in highly sensitive financial environments.

Fourth, multi-agent reinforcement learning frameworks can be explored to simulate collaborative financial planning scenarios involving families or business entities.

Fifth, real-time integration with banking APIs and open finance ecosystems can enable automatic transaction streaming and continuous learning without manual data upload.

Finally, explainable AI mechanisms can be embedded within the conversational layer to provide transparent reasoning behind financial recommendations, thereby increasing user trust and regulatory compliance readiness.

With these advancements, the system can evolve into a comprehensive, autonomous financial intelligence platform capable of supporting long-term financial planning, risk mitigation, and intelligent wealth management in modern digital economies.

7.REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 8, pp. 9452–9468, 2023.

[2] O. Lithgow-Serrano, M. Tan, and J. Kim, “Assessing Retrieval-Augmented Generation Capabilities for Financial Documents,” Proceedings of the Financial NLP Conference, 2025.

[3] M. Rizinski, “AI Agents for Financial Analytics and Advisory: A Survey of Recent Advances,” International Journal of Financial Technology, vol. 9, no. 4, pp. 212–238, 2025.

[4] Y. Cao, Z. Chen, and R. Kumar, “RiskLabs: Predicting Financial Risk Using Large Language Models and Multi-Source Data,” IEEE Access, vol. 12, pp. 45120–45138, 2024.

[5] H. Zhao, Q. Yao, and J. Kwok, “Deep Reinforcement Learning for Sequential Recommendation Systems,” ACM Transactions on Information Systems, vol. 41, no. 3, 2023.

[6] Y. Li and T. Chen, “Reward Shaping Strategies for Health and Financial Sequential Recommendation,” Expert Systems with Applications, vol. 240, 2024.

[7] S. Rendle, “Factorization Machines and Context-Aware Recommendation Models: Recent Advances,” ACM Transactions on

Intelligent Systems and Technology, vol. 14, no. 2, 2023.

[8] L. Wu, X. He, X. Wang, and M. Wang, "Graph Neural Networks in Recommender Systems: A Comprehensive Review," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 1, pp. 1–21, 2023.

[9] H. N. Bhandari, S. Kumar, and R. Sharma, "Deep Learning-Based Predictive Modeling for Financial and Market Volatility Analysis," IEEE Access, vol. 12, pp. 31789–31805, 2024.

[10] J. Wang, K. Liu, and H. Li, "Explainable Artificial Intelligence for Financial Decision Support Systems," Artificial Intelligence in Finance, vol. 6, no. 2, pp. 101–118, 2024.

[11] M. Rostami and K. Berahmand, "Hybrid Machine Learning Models for Financial Fraud Detection," Neurocomputing, vol. 575, pp. 134–149, 2024.

[12] A. Srinivasan et al., "Enhancing Financial Question Answering Using RAG and Agentic LLMs," arXiv preprint arXiv:2509.16369, 2025.

[13] D. Millo, B. Vika, and N. Baci, "Integrating NLP Techniques into Intelligent Financial Information Systems," Information Systems Research Journal, vol. 32, no. 4, pp. 540–558, 2024.

[14] J. Brown and E. Carter, "Vector Databases and Scalable Embedding Retrieval for Enterprise AI Applications," Journal of Big Data Systems, vol. 10, no. 3, pp. 88–104, 2023.

[15] K. Patel and S. Gupta, "Anomaly Detection in Financial Transactions Using Isolation Forest and Transformer Models," Applied Soft Computing, vol. 138, 2024.