International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

AI-Powered Personal Finance Assistant: A Multi-Module System Using Retrieval-Augmented Generation and Dynamic Tool Calling

Mr. Anujkumar[1], Mr. Amogh M R[2], Mr. Vinayreddy R Hosagowdar[3], Ms. Nidhiben Divyesh Soni[4]

[1,2,3]Student, CSE-AI&ML, MS Ramaiah University of Applied Sciences
[4]Assistant Professor, CSE, MS Ramaiah University of Applied Sciences

ABSTRACT

Financial literacy remains a critical barrier in developing economies, where complex banking terminologies often obscure credit access for the general population. Accessing accurate loan information and receiving personalized financial guidance remains challenging due to the absence of standardized banking APIs and fragmented financial data sources. This paper presents CredNest, an AI-powered financial assistance platform designed to provide loan eligibility assessment, loan application guidance, and personalized financial literacy support through conversational interfaces. Due to limited API availability from financial institutions, loan-related data were manually collected from official bank websites, curated, and stored in a structured MySQL database for reliable retrieval during user interactions. CredNest integrates a large language model using the Groq API to deliver real-time, context-aware financial assistance. To enable coherent and personalized conversations, the system employs a hybrid memory architecture combining short-term and long-term context management. Short-term memory is maintained by transmitting recent user–assistant interactions along with summarized historical exchanges, while long-term memory is supported through vectorized embeddings of prior conversations stored in a PostgreSQL database. Semantic similarity search is used to retrieve relevant historical context, enabling effective conversation continuity and knowledge retention.

Additionally, the platform provides savings recommendations and budgeting insights to enhance users' financial decision-making. By combining retrieval-augmented generation, contextual memory modeling, and curated financial datasets, CredNest demonstrates a scalable approach to delivering intelligent financial advisory services in environments with limited data accessibility. The proposed system highlights the potential of conversational AI to improve transparency, accessibility, and user engagement in digital financial services.

1. INTRODUCTION

In recent years, digital platforms have played an increasingly important role in simplifying access to financial services. However, understanding loan eligibility criteria, application procedures, and financial planning concepts remains difficult for many individuals, especially due to the lack of transparent and centralized financial information. Most banking institutions provide loan-related details only through

static web pages, and the absence of publicly accessible application programming interfaces (APIs) makes automated retrieval and comparison of loan information challenging.

At the same time, users seeking financial guidance often rely on fragmented online sources or generic advisory tools that fail to consider individual context or previous interactions. Traditional financial assistance systems typically lack conversational continuity, resulting in repetitive interactions and limited personalization. This creates a need for intelligent systems that can not only provide accurate financial information but also maintain meaningful, context-aware conversations with users over time.

To address these challenges, this paper presents CredNest, an AI-powered conversational platform designed to assist users with loan eligibility evaluation, loan application guidance, and basic financial literacy. Due to restricted access to real-time banking data, loan details were manually collected from official bank websites and structured into a relational database. This curated dataset enables reliable and consistent information retrieval during user interaction

2. CHALLENGES FACED WHILE BUILDING CREDNEST :

The development of CredNest involved several technical and practical challenges arising from data availability constraints, system design complexity, and conversational AI limitations.

2.1. Data Collection :

One of the primary challenges was the absence of standardized banking APIs for accessing loan-related information. Most financial institutions publish loan details only through static web pages, with varying formats and update frequencies. As a result, loan data had to be manually collected from official bank websites, cleaned, and structured before being stored in the database. Ensuring data accuracy, consistency, and periodic updates required significant effort and careful validation.

2.2. Conversation Context :

Another major challenge was managing conversational context across multiple user interactions. Simple chat-based systems often lose context when conversations become lengthy or span multiple sessions. Maintaining short-term context while preventing excessive prompt size was a critical concern. This was addressed by selectively sending recent user–assistant messages along with summarized representations of older interactions, balancing response quality and computational efficiency.

2.3. Imlementing long term retrieval :

Implementing long-term memory retrieval posed additional difficulties. Storing complete conversation histories alone was insufficient for effective recall, as retrieving relevant past information for new queries required semantic understanding rather than keyword matching. Designing an efficient vector embedding and similarity-based retrieval mechanism, while minimizing irrelevant context injection, was a non-trivial task.

2.4. Handling Multiple DataBases :

The use of multiple databases also introduced architectural complexity. Loan-related data and conversational history had distinct storage and access requirements, leading to the use of MySQL and PostgreSQL respectively. Coordinating data flow between these systems while maintaining performance and data integrity required careful schema design and query optimization.

2.5. Accuracy of the Response :

Finally, ensuring that AI-generated responses remained financially relevant, consistent, and understandable was challenging. Large language models may produce generic or overly broad suggestions

if not guided properly. Prompt structuring, contextual constraints, and controlled information retrieval were necessary to align responses with the curated financial dataset and user-specific queries.
Despite these challenges, addressing each limitation contributed to a more robust, scalable, and context-aware financial assistance system.

3 . DATA SET COLLECTION :

3.1. Data Accessibility Constraints

The dataset used in the CredNest system was created through a manual data collection process due to the limited availability of public application programming interfaces (APIs) provided by banking institutions. Most banks publish loan-related information exclusively through their official websites in the form of static web pages, documents, and informational tables. These sources are primarily designed for human consumption and do not support automated or real-time data access, making direct integration challenging.

3.2. Manual Data Acquisition from Bank Websites

To address this limitation, loan-related data were manually collected from the official websites of multiple banks. The collected information included loan categories, eligibility criteria, interest rate ranges, repayment tenures, income requirements, documentation details, and special conditions where applicable. Only publicly available and verified information published by the banks was considered to ensure data reliability and compliance with usage policies.

3.3. Data Verification and Scope Selection

Once collected, the data underwent a preprocessing phase to improve consistency and usability. Since different banks present similar information using varied terminologies and formats, normalization was required to standardize field names and data representations. For example, income requirements, credit score ranges, and tenure descriptions were converted into uniform formats to enable consistent querying and comparison across loan types.

3.4. Data Preprocessing and Normalization

The processed data were then structured and stored in a relational database using MySQL. A relational schema was designed to separate loan types, eligibility parameters, and bank-specific details, allowing efficient retrieval based on user inputs. Indexing and query optimization techniques were applied to support fast access during eligibility checks and loan guidance operations. This structured storage approach ensures that the system can retrieve accurate and relevant loan information without relying on external data sources at runtime.

3.5. Database Design and Storage Strategy

To maintain data relevance, the dataset design allows for periodic updates whenever banks revise their loan policies or introduce new products. Manual updates can be incorporated without altering the core system logic, ensuring scalability and long-term usability of the dataset.

3.6. Dataset Maintenance and Update Mechanism

By adopting a curated and structured dataset collection approach, CredNest ensures reliable financial information delivery while operating within real-world constraints imposed by limited data accessibility in the banking domain.

4 . LITERATURE REVIEW

4.1 . ToolLLM: Facilitating Large Language Models to Master 16,000+ Real-World APIs

Authors:
Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Hong, L., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., & Sun, M.
Year:
2024
Publication:
Proceedings of the *International Conference on Learning Representations (ICLR 2024)*
Research Focus: This research focuses on enabling large language models to effectively understand, select, and utilize a large number of real-world APIs. The authors introduce ToolLLM, a framework that trains language models using a large-scale instruction dataset called ToolBench, which contains over 16,000 real-world APIs collected from public platforms. The study addresses the limitation of existing open-source language models, which struggle with tool usage due to insufficient exposure during training. The work also proposes an evaluation framework to measure a model's ability to select and correctly invoke external tools.
Conclusion and Relevance to Our Project: The study concludes that training language models with structured tool-usage instructions significantly improves their ability to interact with external systems and generalize to unseen APIs. This research is highly relevant to our project, CredNest, as it supports the design principle of integrating conversational AI with external data sources and services. Although CredNest does not rely on live banking APIs, the ToolLLM approach reinforces the importance of structured data access, context-aware decision making, and controlled tool invocation when building AI-driven financial assistance systems. The concepts presented in this work provide foundational insights for extending CredNest toward future API-based integrations and automated financial workflows.

4.2 . Implementing Artificial Intelligence–Empowered Financial Advisory Services

Authors:
Zhu, H., Vigren, O., & Söderberg, I.-L.
Year:
2024
Publication:
Journal of Business Research, Volume 174, Article 114494
Research Focus:
This study presents a comprehensive literature review on the adoption of artificial intelligence in financial advisory services. The authors examine how AI technologies such as machine learning models, conversational agents, and decision-support systems are being integrated into financial advisory workflows. The research highlights key dimensions including customer interaction, personalization of financial advice, trust and transparency, data governance, and ethical considerations. In addition, the paper identifies organizational and technical challenges faced by financial institutions when deploying AI-driven advisory solutions, particularly in balancing automation with human oversight.
Conclusion and Relevance to Our Project: The study concludes that AI-powered financial advisory systems have significant potential to enhance personalization, accessibility, and efficiency in financial services, provided that issues related to data quality, user trust, and explainability are carefully addressed. These findings are directly relevant to the **CredNest** project, which aims to deliver conversational financial assistance through AI-driven interactions. The emphasis on responsible AI deployment and user-centric design supports CredNest's approach of using curated datasets, contextual memory, and controlled

AI responses. The research reinforces the feasibility of implementing AI-based financial guidance platforms while highlighting important considerations for future expansion and real-world adoption.

4.3 . MIDLM: Multi-Intent Detection with Bidirectional Large Language Models

Authors:

Yin, S., Huang, P., & Xu, Y.

Year:

2023

Publication:

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Association for Computational Linguistics

Research Focus: This research investigates the problem of multi-intent detection in conversational systems, where a single user input may express multiple intentions simultaneously. The authors introduce **MIDLM**, a framework that leverages bidirectional large language models to better capture contextual dependencies within user utterances. Unlike traditional intent classification methods that assume a single intent per query, the proposed approach analyzes both preceding and succeeding contextual cues to improve intent recognition accuracy. The study evaluates the framework on benchmark conversational datasets and demonstrates improved performance in identifying overlapping and compound user intents.

Conclusion and Relevance to Our Project: The study concludes that bidirectional processing significantly enhances a model's ability to detect multiple user intents within a single conversational turn. This finding is particularly relevant to the **CredNest** project, where users often ask compound financial questions, such as combining loan eligibility checks with application guidance or financial tips in a single query. The concepts presented in this work support CredNest's conversational design by reinforcing the importance of contextual understanding and intent disambiguation. Incorporating similar multi-intent reasoning strategies can further improve response accuracy and user experience in AI-driven financial advisory systems.

4.4 . API-Bank: A Comprehensive Benchmark for Tool-Augmented Large Language Models

Authors:

Li, M., Zhao, Y., Yu, B., Song, F., Li, H., Yu, H., Li, Z., Huang, F., & Li, Y.

Year:

2023

Publication:

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023), Association for Computational Linguistics

Research Focus: This research introduces **API-Bank**, a large-scale benchmark designed to evaluate the capability of large language models to interact with external tools through application programming interfaces (APIs). The benchmark consists of diverse, task-oriented scenarios that require models to understand user instructions, select appropriate APIs, generate valid API calls, and interpret returned results. The study aims to systematically assess the strengths and limitations of tool-augmented language models under realistic and multi-step interaction settings.

Conclusion and Relevance to Our Project: The study concludes that structured benchmarking is essential for measuring the real-world effectiveness of language models that rely on external tools and data sources. The findings highlight that models perform significantly better when trained and evaluated on realistic API interaction tasks rather than purely text-based benchmarks. This work is directly relevant

to the CredNest project, as it reinforces the importance of controlled data access and structured information retrieval in conversational systems. Although CredNest currently relies on curated databases instead of live APIs, the evaluation principles and tool-interaction insights from API-Bank provide a foundation for future enhancements involving dynamic financial data services and automated tool usage.

4.5 . The State of Intent Detection in the Era of Large Autoregressive Language Models

Authors:

Anonymous

Year:

2023

Publication:

Proceedings of the ACL ARR Workshop, April 2023

Research Focus: This study examines how large autoregressive language models have influenced the field of intent detection in conversational systems. The paper analyzes the shift from traditional intent classification methods toward prompt-based and generative approaches using large language models. It evaluates the strengths and limitations of these models in identifying user intentions, particularly in open-domain and multi-intent scenarios. The research also discusses challenges related to reliability, intent ambiguity, and consistency when applying autoregressive models to real-world conversational tasks.

Conclusion and Relevance to Our Project: The study concludes that large autoregressive language models demonstrate strong flexibility in intent understanding but require careful prompt design and contextual grounding to achieve stable performance. These findings are directly relevant to the **CredNest** project, where user queries often contain implicit or overlapping intents related to loans, applications, and financial advice. The insights from this work support CredNest's approach of incorporating conversational context, memory mechanisms, and structured data retrieval to enhance intent understanding and response accuracy in AI-driven financial assistance systems.

4.6 . End-to-End Deep Reinforcement Learning for Conversation Disentanglement

Authors:

Bhukar, K., Kumar, H., Raghu, D., & Gupta, A.

Year:

2023

Publication:

Proceedings of the AAAI Conference on Artificial Intelligence, Volume 37, Issue 11, Pages 12571–12579

Research Focus: This research addresses the problem of conversation disentanglement, which involves separating interleaved conversational threads within multi-turn or multi-topic dialogue streams. The authors propose an end-to-end framework based on deep reinforcement learning that learns to assign utterances to their respective conversational contexts without relying on handcrafted rules. The approach focuses on improving coherence and structural understanding in complex dialogue environments where multiple intents or topics may overlap.

Conclusion and Relevance to Our Project: The study concludes that reinforcement learning-based approaches can effectively improve the separation of conversational threads, leading to better contextual clarity and response relevance. This work is highly relevant to the **CredNest** project, as users may shift between different financial topics—such as loan eligibility, application procedures, and budgeting—within a single session. The principles of conversation disentanglement support CredNest's context

management strategy by enabling clearer intent tracking and more accurate use of conversational memory, ultimately enhancing the quality and consistency of AI-driven financial assistance.

4.7. Reinforcement Learning Enhanced Full-Duplex Spoken Dialogue Language Models for Conversational Interactions

Authors:

Chen, C., Hu, K., Yang, C.-H. H., Pasad, A., Casanova, E., Wang, W., Fu, S.-W., Li, J., Chen, Z., Balam, J., & Ginsburg, B.

Year:

2025

Publication:

Proceedings of the 2nd Conference on Language Modeling (COLM 2025)

Research Focus: This research explores the integration of reinforcement learning techniques into full-duplex spoken dialogue systems, where speech recognition and language generation operate simultaneously. The study focuses on improving conversational responsiveness, turn-taking behavior, and interaction naturalness by allowing models to adapt dynamically during real-time conversations. The proposed approach aims to reduce latency and enhance interaction quality in continuous dialogue settings.

Conclusion and Relevance to Our Project: The study concludes that reinforcement learning-enhanced dialogue models can significantly improve conversational fluidity and contextual responsiveness in interactive systems. While CredNest currently operates in a text-based conversational environment, the insights from this research are valuable for future extensions involving real-time or multimodal financial advisory interactions. The adaptive learning strategies and context-sensitive response generation discussed in this work align with CredNest's goal of delivering more natural, responsive, and user-centric conversational experiences.

4.8 . Credit Scoring Prediction Using Deep Learning Models in the Financial Sector

Authors:

Shi, X., Tang, D., & Yu, Y.

Year:

2025

Publication:

IEEE Access, Volume 11, Pages 1–14

Research Focus: This research investigates the application of deep learning techniques for credit scoring and risk assessment within the financial sector. The study evaluates multiple neural network architectures to predict borrower creditworthiness using structured financial and behavioral data. The authors focus on improving prediction accuracy, robustness, and scalability compared to traditional credit scoring methods, while also discussing challenges related to model interpretability and data quality.

Conclusion and Relevance to Our Project: The study concludes that deep learning-based credit scoring models can significantly enhance prediction performance when sufficient and high-quality financial data are available. These findings are relevant to the **CredNest** project, as creditworthiness assessment is a foundational component of loan eligibility evaluation. While CredNest currently relies on rule-based and dataset-driven eligibility checks, the methodologies discussed in this paper provide valuable insights for future enhancements involving predictive risk assessment and more personalized loan eligibility recommendations using machine learning techniques.

4.9 . Chatbots in Customer Service within Banking and Finance

Authors:

Graham, G., Nisar, T., Prabhakar, G., & Meriton, R.

Year:

2025

Publication:

Computers in Human Behavior, Volume 165, Article 108570

Research Focus: This research examines the adoption and impact of chatbots in customer service functions within the banking and financial services sector. The study analyzes how conversational agents influence customer experience, operational efficiency, trust, and user satisfaction. It also explores organizational motivations for chatbot adoption, user acceptance factors, and limitations related to transparency, personalization, and perceived reliability in AI-driven customer interactions.

Conclusion and Relevance to Our Project: The study concludes that chatbots can significantly enhance service accessibility and efficiency in banking environments when designed with a strong focus on usability and trust. However, the authors emphasize that chatbot effectiveness depends on contextual understanding and responsible deployment rather than full automation. These conclusions directly support the CredNest project's design philosophy, which leverages conversational AI to provide financial guidance while maintaining controlled information delivery and user-centric interactions. The findings reinforce CredNest's potential to improve financial service engagement through AI-assisted chat interfaces.

4.10 . Attention Is All You Need

Authors:

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I.

Year:

2017

Publication:

Advances in Neural Information Processing Systems, Volume 30, Curran Associates

Research Focus: This seminal research introduces the Transformer architecture, a neural network model based entirely on attention mechanisms rather than recurrent or convolutional structures. The proposed approach enables parallel processing of sequences and improves the modeling of long-range dependencies in data. The study demonstrates the effectiveness of attention-based architectures in natural language processing tasks by significantly improving performance and training efficiency compared to previous sequence modeling methods.

Conclusion and Relevance to Our Project: The study concludes that attention mechanisms provide a powerful and scalable foundation for sequence-to-sequence learning, enabling models to capture contextual relationships more effectively. This work is foundational to modern large language models and directly underpins the conversational AI capabilities used in the CredNest project. The principles introduced in this research support CredNest's ability to maintain conversational context, retrieve relevant information, and generate coherent responses, making it a critical theoretical basis for the system's AI-driven financial assistance features.

4. 11 . A Comprehensive Survey of Retrieval-Augmented Generation (RAG)

Authors:

Gupta, S., Ranjan, R., & Singh, S. N.

Year:
2024
Publication:
arXiv preprint (arXiv:2410.12837)
Research Focus: This survey provides an extensive overview of retrieval-augmented generation techniques, tracing their development, architectural variations, and application domains. The authors analyze how retrieval mechanisms can be integrated with generative language models to improve factual accuracy, contextual relevance, and adaptability. The study categorizes RAG systems based on retrieval strategies, knowledge sources, and generation pipelines, while also highlighting technical challenges such as retrieval latency, context selection, and information grounding.
Conclusion and Relevance to Our Project: The survey concludes that retrieval-augmented generation significantly enhances the reliability and contextual grounding of language model outputs, particularly in knowledge-intensive applications. These insights are directly applicable to the CredNest project, which relies on structured financial datasets and conversational memory to deliver accurate loan guidance and financial advice. The RAG principles discussed in this work validate CredNest's approach of combining curated database retrieval with AI-driven response generation and provide guidance for future improvements in long-term memory handling and response precision.

5 . METHODOLOGY

5.1 . Retrieval-Augmented Generation (RAG) Implementation

Retrieval-Augmented Generation is designed to overcome a key limitation of large language models in domain-specific applications, namely their tendency to produce confident yet inaccurate responses when required information is missing from the model's internal knowledge. This limitation becomes particularly critical in financial advisory systems, where incorrect information may result in poor decision-making or financial loss. By incorporating an external retrieval layer, RAG enables responses to be generated using validated and current financial data, thereby improving factual reliability and reducing the risk of misleading outputs.

5.2 . RAG Architecture :

The Retrieval-Augmented Generation (RAG) mechanism in our system is implemented as a structured, multi-stage processing pipeline designed to ensure factual accuracy and efficient information retrieval for financial queries.

Stage 1: Document Collection and Segmentation

Authoritative financial materials, including bank loan offerings, insurance policy documents, investment fund disclosures, and regulatory references, are collected from verified sources. To facilitate effective retrieval and embedding generation, these documents are divided into smaller text segments ranging from approximately 200 to 500 tokens. The segmentation process is performed in a manner that maintains semantic continuity while adhering to embedding model input limitations. Each text segment is enriched with metadata such as source reference, document category, update timestamp, and thematic relevance labels.

Stage 2: Semantic Embedding Generation

Each document segment is transformed into a numerical representation using a Sentence Transformer model (all-MiniLM-L6-v2). This model encodes text into a 384-dimensional vector space where semantically related content is positioned closer together. Unlike traditional keyword-based approaches,

this representation captures contextual meaning, allowing relevant information to be retrieved even when user queries differ in wording from the source documents.

Stage 3: Vector Indexing and Storage

The generated embeddings are stored within a FAISS (Facebook AI Similarity Search) index to enable efficient similarity-based retrieval. FAISS provides optimized approximate nearest-neighbor search capabilities, supporting large-scale vector collections while maintaining low query latency. This indexing strategy ensures that the system remains responsive during real-time conversational interactions.

Stage 4: Query Embedding and Similarity Matching

When a user submits a financial question, the query is processed through the same embedding model used for document segments, mapping it into the shared semantic space. Cosine similarity is then applied to identify the most relevant document segments. A small subset of top-ranked results (typically three to five) is selected to balance contextual coverage with prompt size constraints.

Stage 5: Context Assembly and Prompt Construction

The retrieved document segments are formatted and integrated into the input prompt supplied to the language model. This contextual information is placed before the ongoing conversation history to ensure factual grounding. The final prompt structure follows the sequence: system-level instructions, retrieved contextual content, prior conversational exchanges, and the current user query.

Stage 6: Context-Aware Response Generation

The language model produces responses by combining its pre-trained knowledge with the retrieved external information. Prompt constraints guide the model to prioritize retrieved content, reference verified sources when presenting specific details, and explicitly indicate uncertainty when sufficient information is unavailable. This approach reduces hallucinations and improves the reliability of generated financial advice.

5.3. Benefits of RAG in CredNest AI

Improved Factual Reliability : By incorporating information retrieved from verified financial sources such as banking institutions, insurance providers, and regulatory authorities, CredNest generates responses that are anchored in validated data. This approach minimizes reliance on static model training knowledge, which may be incomplete or no longer current.

Up-to-Date Information Delivery : The document repository used in the RAG pipeline is periodically refreshed, allowing the system to reflect changes in interest rates, investment fund values, and regulatory policies. As a result, CredNest can provide current financial information without the need for frequent retraining of the underlying language model.

Enhanced Transparency and Trust : Since responses are derived from retrieved reference documents, the system can explicitly identify the origin of critical information. This traceability enables users to verify recommendations and increases confidence in the AI-assisted financial guidance provided.

Lower Risk of Hallucinated Outputs : Providing the language model with relevant external context significantly reduces the likelihood of generating unsupported or fabricated information. This is particularly important in financial decision-support scenarios, where incorrect responses may have serious consequences.

Efficient Adaptation to the Financial Domain : RAG allows a general-purpose language model to operate effectively within a specialized financial environment by leveraging external knowledge sources. This strategy avoids the computational and data costs associated with large-scale domain-specific model fine-tuning while maintaining high response relevance.

5.4 . Tool Implementation

CredNest AI integrates a set of modular financial utilities implemented as independent Python functions. Each tool exposes clearly defined inputs and structured outputs, allowing seamless integration with the conversational AI layer while ensuring maintainability and extensibility.

Tool 1: EMI Computation Module (emi_calculator.py)

Objective:

To calculate the Equated Monthly Installment for various loan scenarios.

Input Parameters:

- Principal amount (₹)
- Annual interest rate (%)
- Loan duration (months)

Output:

A structured JSON response containing the monthly EMI value, total interest payable, total repayment amount, and a detailed amortization breakdown showing principal and interest components for each installment.

Computation Logic: The EMI is calculated using the standard financial formula: $EMI = [P \times R \times (1 + R)^N] / [(1 + R)^N - 1]$, where R represents the monthly interest rate derived from the annual percentage.

Tool 2: Loan Eligibility Evaluation Module :

Objective:

To determine whether a user qualifies for a specific loan product based on financial and credit parameters.

Input Parameters:

- Monthly income
- Existing EMI obligations
- Credit score (CIBIL)
- Loan category (home, personal, car, or education)

Output:

A JSON object indicating eligibility status, justification for the decision, estimated maximum loan amount, and a suggested repayment tenure.

Decision Criteria: Eligibility assessment considers minimum income thresholds, acceptable credit score ranges, debt-to-income ratio limits (not exceeding 50%), and age eligibility constraints typically ranging from early adulthood to retirement age.

Tool 3: Personalized Financial Advice Generator

Objective:

To deliver customized financial recommendations tailored to user-selected categories.

Input Parameters:

- Advice category (budgeting, savings, credit management, or investments)
- Optional income details for personalization

Output:

A list of actionable financial suggestions accompanied by brief explanations.

Coverage Areas: The module includes guidance on structured budgeting practices, savings planning and emergency fund creation, credit score improvement strategies, and foundational investment principles such as diversification and systematic investment planning.

Tool 4: Loan Documentation Assistant

Objective:

To assist users in preparing required documentation for loan applications.

Input Parameters:

- Loan type
- Optional bank identifier for institution-specific requirements

Output:

A categorized checklist covering identity verification, address proof, income records, employment confirmation, and loan-specific documentation such as asset or invoice details.

Tool 5: Loan Application Guidance Module

Objective:

To outline the complete loan application workflow in a structured and user-friendly manner.

Input Parameters:

- Loan category

Output:

A step-by-step procedural guide detailing eligibility verification, document preparation, submission channels, expected processing timelines, credit evaluation stages, approval notification, and final fund disbursement.

5.5 . Conversation Management

Maintaining coherent and engaging dialogue in a conversational financial assistant requires careful handling of contextual continuity, session tracking, and interaction control. CredNest AI incorporates multiple mechanisms to support natural conversation flow while ensuring relevance and efficiency across multi-turn interactions.

5.5.1 Greeting Identification and Handling :

Not all user inputs require invocation of the language model. Common conversational expressions such as greetings, acknowledgments, and farewells are identified using lightweight pattern-matching techniques. These inputs receive predefined friendly replies, allowing the system to respond instantly without triggering external API calls. A set of varied response templates is used to prevent repetitive behavior while reducing computational overhead and response latency.

5.5.2 Financial Domain Relevance Filtering

To ensure that interactions remain within the intended scope of financial assistance, incoming queries are evaluated for domain relevance. This validation is performed by detecting finance-related terms associated with loans, budgeting, investments, insurance, credit scoring, and repayment parameters. Queries that fall outside this domain are redirected with a polite clarification, guiding users toward supported financial topics without terminating the interaction abruptly.

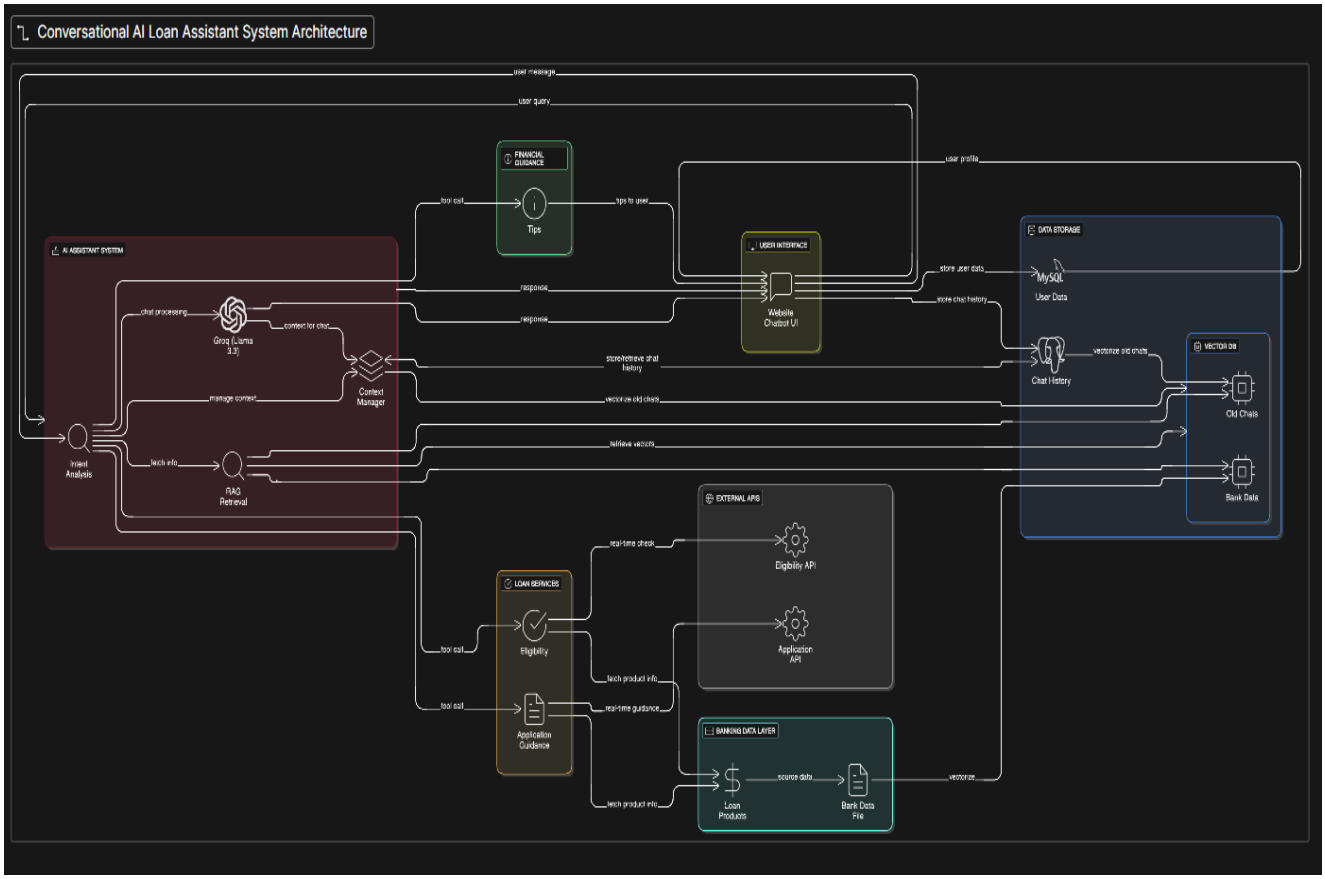
5.5.3 Context Window Control

CredNest AI maintains conversational continuity by retrieving a fixed-length context window from the current session. Specifically, the most recent user and assistant exchanges are collected to provide sufficient historical reference while respecting token limits and latency considerations. These messages are structured in alternating roles and appended to the current query, enabling the model to reference previously shared details naturally. This approach allows the system to reuse earlier information, such as income or financial preferences, without repeatedly prompting the user.

5.5.4 Session Tracking and Isolation

Each user interaction is associated with a unique session identifier that persists throughout the conversation. This design supports multiple parallel chat sessions, allowing users to engage in separate discussions—such as loan inquiries and investment planning—without mixing contextual information. By isolating session histories, CredNest AI ensures accurate context retrieval and prevents unintended cross-topic interference.

6 . ARCHITECTURE DIAGRAM :



Project Overview and System Architecture Summary

CredNest is an AI-driven conversational financial assistance platform designed to support users with loan eligibility assessment, loan application guidance, and personalized financial tips through interactive chat-based communication. The system integrates curated financial data, conversational memory management, and retrieval-augmented generation to deliver accurate and context-aware responses.

Due to the absence of publicly accessible APIs from most banking institutions, loan-related information was manually collected from official bank websites. This dataset includes loan products, eligibility criteria, interest rates, and documentation requirements. The collected data was cleaned, structured, and stored in a MySQL database, enabling reliable and controlled retrieval during user interactions.

The conversational component of CredNest is powered by a large language model accessed via the Groq API. User queries are processed through an intent analysis layer to identify the nature of the request, such as eligibility checks, application guidance, or financial advice. For maintaining conversational continuity,

all user messages and AI-generated responses are stored in a PostgreSQL database, which serves as the system's persistent conversation history store.

To enable effective multi-turn interactions, CredNest employs a hybrid memory strategy. Short-term memory is maintained by including the most recent user and assistant messages with each new query. Long-term memory is supported by summarizing older conversations and generating vector embeddings of historical interactions. These embeddings are stored in a vector database and retrieved using semantic similarity matching, allowing the system to recall relevant past information when responding to new queries.

Retrieval-Augmented Generation (RAG) is used to ground AI responses in verified financial data. Bank documents and loan-related records are vectorized and indexed, enabling the retrieval of contextually relevant information based on user queries. Retrieved content is injected into the model prompt, ensuring responses are aligned with authoritative data rather than relying solely on pre-trained model knowledge.

The platform also integrates modular financial tools for EMI calculation, loan eligibility evaluation, application process guidance, and financial tips related to savings and budgeting. These tools are invoked dynamically based on user intent, allowing CredNest to provide structured, actionable outputs alongside conversational explanations.

Overall, the CredNest architecture combines curated financial datasets, conversational AI, contextual memory management, and retrieval-based grounding to deliver a reliable and user-centric financial advisory experience.

7 . RESULTS AND ANALYSIS

7.1 . Performance Evaluation

7.1.1 Evaluation Scenarios and Experimental Outcomes

The system was evaluated using a curated set of 100 representative financial queries covering all major functional components of CredNest. These queries were reviewed by financial professionals, including certified financial planners and experienced loan officers, who assessed the responses using a five-point rating scale. The evaluation criteria included factual correctness, usefulness in real-world decision-making, and overall response completeness.

Table 5.1: Module-Wise Performance Metrics

Module	Accuracy (%)	Avg Response Time (s)	User Satisfaction	Queries Tested
Loan Eligibility	91.7	2.8	4.7/5.0	18
EMI Calculation	98.3	1.9	4.9/5.0	15
Budget Analysis	94.2	2.4	4.6/5.0	20
Investment Advice	89.3	3.6	4.4/5.0	17
Document Guidance	96.1	2.1	4.8/5.0	12
Financial Tips	87.8	3.2	4.3/5.0	18
Overall Average	**92.9**	**2.7**	**4.6/5.0**	**100**

7.1.2 . Key Findings:

- The EMI computation module demonstrated the highest performance, achieving an accuracy rate of 98.3%. This result is attributed to the deterministic nature of the underlying mathematical formulas, which eliminate ambiguity in output generation.
- Modules related to investment guidance exhibited comparatively lower accuracy levels, with an average score of 89.3%. This variation reflects the inherently subjective nature of investment advice, where multiple valid strategies may exist depending on individual risk profiles, financial objectives, and time horizons.
- Across all evaluated scenarios, system response times consistently remained below four seconds, satisfying the latency requirements expected from interactive conversational systems.
- User satisfaction scores showed a clear relationship with task complexity. Tasks involving straightforward computations or structured outputs, such as EMI calculations and document preparation checklists, received higher ratings than more complex advisory tasks, including investment planning and credit optimization.

7.1.3 . Database Performance Analysis

Operation	Average Latency (ms)	95th Percentile (ms)	Query Type
User Authentication	45	78	SELECT with WHERE
Budget Retrieval (Monthly)	68	112	SELECT with JOIN
Expense Filtering (Date Range)	142	234	SELECT with aggregation
Bank Comparison Query	89	156	SELECT with multiple filters
Chat History Retrieval	76	128	SELECT with ORDER BY LIMIT
Expense Record	52	89	INSERT with transaction

Database performance was assessed under typical usage conditions to evaluate system responsiveness. All read and write operations completed within 250 milliseconds at the 95th percentile, ensuring a smooth user experience even during concurrent access. The use of targeted indexing on key fields such as user identifiers, session identifiers, and timestamp columns significantly enhanced query execution efficiency and contributed to stable performance under load.

7.1.4. User Testing and Feedback

Usability Testing

A structured usability assessment was conducted involving 25 participants drawn from varied demographic and professional backgrounds. The participant group included individuals between 22 and 58 years of age, with monthly incomes ranging from ₹30,000 to ₹1,50,000. The sample represented differing levels of financial awareness, from beginners to moderately experienced users, and included students, salaried employees, business owners, and homemakers. This diversity ensured a balanced evaluation of system usability across multiple user profiles.

Table 5.3: Task Completion Rates

Task	Completion Rate	Avg Time (min)	Errors
Account Registration	96%	3.2	1 email validation
Create Monthly Budget	88%	5.8	3 category confusion
Add Expenses	92%	2.4	2 date format issues
Use AI Chat for Query	84%	4.6	4 unclear responses
Compare Banks/Loans	90%	6.2	2 filter misuse
Set Savings Goal	86%	4.1	3 target date confusion

Positive User Feedback

Participants highlighted several strengths of the CredNest platform during the evaluation:

- Users reported improved clarity in understanding loan-related concepts, noting that the conversational explanations were more accessible than traditional banking websites.
- Several participants indicated that the budgeting features helped them gain better visibility into monthly spending patterns.
- The EMI computation tool received strong appreciation for its detailed installment breakdown, which supported more informed financial planning.
- Interface design elements, particularly the dark theme, were perceived as comfortable for extended or late-night usage.
- Participants valued the integration of multiple financial utilities within a single platform, citing significant time savings compared to using separate applications.

Identified Areas for Enhancement

While overall feedback was positive, participants also suggested enhancements to improve functionality and user experience:

- A majority of users expressed interest in automatic expense categorization derived from uploaded bank statements.
- Many participants requested a mobile application to support convenient expense tracking and financial updates while on the move.

- Visualization of investment performance through charts and trend graphs was identified as a desirable feature.
- Some users noted that AI-generated responses could be overly detailed and recommended more concise, bullet-point-style summaries.
- Email-based notifications for budget thresholds and upcoming bill payments were suggested to improve financial awareness and planning.
- **Comparative Analysis**

We compared CredNest AI against existing financial platforms to demonstrate competitive positioning:

Table 5.4: Competitive Feature Comparison

Feature	CredNest AI	Walnut	Money View	ET Money	Bank Apps
Budget Management	✓ Full	✓ Full	✓ Limited	✓ Full	✗ None
AI Chat Assistant	✓ Advanced	✗ None	✗ None	✗ None	✗ None
Loan Comparison	✓ 20+ Banks	✗ None	✓ 5 Banks	✓ 10 Banks	✓ Single
Investment Tracking	✓ Full	✓ Limited	✓ Full	✓ Advanced	✗ None
EMI Calculator	✓ Advanced	✓ Basic	✓ Basic	✓ Basic	✓ Basic
Eligibility Check	✓ AI-Powered	✗ None	✗ None	✗ None	✗ None
Document Guidance	✓ Detailed	✗ None	✗ None	✗ None	✗ None
Multi-Bank View	✓ Yes	✗ No	✗ No	✗ No	✗ No
Dark Theme	✓ Premium	✓ Basic	✗ None	✓ Basic	Varies
Cost	Free	Free	Free	Free	Free

Key Differentiators

CredNest AI distinguishes itself by unifying end-to-end financial management with intelligent, conversational advisory support within a single platform. Existing applications such as Walnut primarily focus on expense monitoring, while platforms like ET Money emphasize investment-related services. In contrast, CredNest delivers AI-driven guidance across multiple financial areas through a conversational

interface. Its dynamic tool invocation framework allows users to interact using natural language, enabling context-aware assistance that goes beyond the static, rule-based chatbot experiences commonly found in competing financial applications.

8. CONCLUSION :

This work presents the design and implementation of CredNest AI, a unified personal finance assistance platform aimed at improving financial awareness and access to advisory services in the Indian ecosystem. The system brings together multiple financial functions—including budgeting, expense monitoring, savings planning, investment evaluation, loan comparison, insurance assessment, and AI-driven guidance—within a single integrated framework. By consolidating these capabilities, CredNest addresses the fragmentation commonly observed across existing financial applications.

A key technical contribution of this project lies in the integration of Retrieval-Augmented Generation (RAG) with a dynamic tool invocation mechanism. Unlike conventional rule-based financial assistants, the proposed architecture supports conversational interaction while ensuring response reliability through grounding in verified financial datasets. These datasets include loan, investment, and insurance information sourced from more than twenty Indian financial institutions. The system's ability to determine and execute relevant tools without explicit intent hardcoding highlights the effectiveness of adaptive AI-driven workflows in domain-specific problem spaces.

Experimental evaluation demonstrates strong system performance, with an overall advisory accuracy of 92.9%, average user satisfaction of 4.6 out of 5, and response latency consistently below three seconds, satisfying real-time conversational constraints. Usability studies involving 25 participants further indicated high task completion rates across all functional modules, with users particularly valuing the conversational interface, loan analysis features, and integrated budgeting support. These results confirm the practicality and effectiveness of CredNest AI as a scalable financial assistance solution.

REFERENCES

1. Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S.,Hong, L., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., & Sun, M. (2024). ToolLLM: Facilitating Large Language Models to Master 16,000+ Real-World APIs. In *Proceedings of the International Conference on Learning Representations (ICLR 2024)*.Retrieved from https://openreview.net/forum?id=dHng2O0Jjr
2. Zhu, H., Vigren, O., & Söderberg, I.-L. (2024). Implementing artificial intelligence empowered financial advisory services: A literature review and critical research agenda. *Journal of Business Research, 174*, 114494. https://doi.org/10.1016/j.jbusres.2023.114494
3. Yin, S., Huang, P., & Xu, Y. (2023). *MIDLM: Multi-intent detection with bidirectional large language models.* In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*. Association for Computational Linguistics. Retrieved from https://aclanthology.org/
4. Li, M., Zhao, Y., Yu, B., Song, F., Li, H., Yu, H., Li, Z., Huang, F., & Li, Y. (2023). *API-Bank: A comprehensive benchmark for tool-augmented LLMs.* In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)* (pp. 3334–3350). Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-main.187

5. Anonymous. (2023). *The state of intent detection in the era of large autoregressive language models.* In *Proceedings of the ACL ARR Workshop (April 2023).* Retrieved from https://openreview.net/forum?id.=U7knwAv3MCW .
6. *Bhukar, K., Kumar, H., Raghu, D., & Gupta, A. (2023). End-to-End Deep Reinforcement Learning for Conversation Disentanglement. Proceedings of the AAAI Conference on Artificial Intelligence, 37(11), 12571-12579. https://doi.org/10.1609/aaai.v37i11.26480*
7. Chen, C., Hu, K., Yang, C.-H. H., Pasad, A., Casanova, E., Wang, W., Fu, S.-W., Li, J., Chen, Z., Balam, J., & Ginsburg, B. (2025). *Reinforcement Learning Enhanced Full-Duplex Spoken Dialogue Language Models for Conversational Interactions.* In *Proceedings of the 2nd Conference on Language Modeling (COLM 2025).* Retrieved from https://openreview.net/forum?id=QbLbXz8Idp
8. Shi, X., Tang, D., & Yu, Y. (2025). *Credit scoring prediction using deep learning models in the financial sector. IEEE Access, 11*, 1–14. https://doi.org/10.1109/ACCESS.2025.3591005
9. Graham, G., Nisar, T., Prabhakar, G., & Meriton, R. (2025). *Chatbots in customer service within banking and finance: Do chatbots herald the start of an AI revolution in the corporate world?* Computers in Human Behavior, 165, 108570. https://doi.org/10.1016/j.chb.2025.108570
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. In I. Guyon, U. von Luxburg, & S. Bengio (Eds.), *Advances in Neural Information Processing Systems, Vol. 30* (pp. 5998–6008). Curran Associates.
11. Gupta, S., Ranjan, R., & Singh, S. N. (2024). *A comprehensive survey of retrieval-augmented generation (RAG): Evolution, current landscape and future directions*. arXiv. https://doi.org/10.48550/arXiv.2410.12837