IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

Privacy-Preserving Offline Financial Assistant with RAG-Based Chatbot and AI Talking Head

Sanskruti Kshirsagar, Shravanee Ramdasi, Eshita Shirolkar, Shubham Bhandwalkar, Vikas Kadam
Department of Information Technology
Marathwada Mitra Mandal's College of Engineering, Pune, Maharashtra, India
kshirsagarsanskruti.02@gmail.com, shravaneeramdasi2254@gmail.com, shirolkareshita@gmail.com,
shubhambhandwalkar2964@gmail.com, vikaskadam@mmcoe.edu.in *(Corresponding author)*

Abstract: *A fresh take begins here. Built right into your gadget, this money helper works solo — no web needed. It thinks fast, acts smart, keeps things locked down tight. Connection? Not required. Cloud storage? Left out on purpose. Safety stays high while speed never drops. The whole system runs where you carry it, always ready, always private.*

A single space holds everything needed: inside it, a local language processor runs alongside a moving face that speaks. This setup allows talk to flow easily, driven by smart software you keep on your own machine. Instead of relying on outside servers, all parts work together in one private loop. The voice and expressions come alive through code built for back-and-forth chat. Everything stays contained, yet behaves like a real exchange. Right inside the machine, everything gets processed. That means personal information stays safer. Users stay in control without relying on outside connections. Even when internet access drops, it keeps working just fine. Security takes priority here, especially where risks run high..

Keywords: Fintech industry, cyber threats, credit card transaction dataset, CNN model, Machine learning

I. INTRODUCTION

New progress in intelligence has led to smart money helpers that guide spending, study finances while talking naturally with people. Still, most current tools depend strongly on servers and constant web access. Raising questions about private data safety, who truly controls the artificial intelligence, whether systems work when needed, and exposure to outside service failures. Worries grow sharper once personal budget details enter the picture.

One way to tackle these problems is to build an artificial intelligence helper for money matters with no internet needed. This version of the intelligence runs only on your own gadget so there is no need for remote servers or constant connections. Because everything happens right where you use the intelligence, your information stays private under your control, leaving no trace online. Instead of sending requests far away, the artificial intelligence tool handles language understand-ing, spending logs, number crunching, finding past entries by meaning, hearing questions, and speaking answers — all inside the machine.

Stored knowledge sits in a structure built into the device, keeping summaries of finances ready for quick clever replies without going online ever. When offline, responses from the artificial intelligence feel sharp, aware, and tied closely to what you asked moments before. Talking happens alongside typing when people log, sort, or review spending over days, weeks, or months with the intelligence. A face appears on screen, speaking with timed mouth shapes and emotions matching words generated on the device by the artificial intelligence. The artificial intelligence works smoothly on everyday laptops or phones with no extra gear needed. Passwords checked at the source keep data private while encryption guards stored files. The artificial intelligence meets conversation inside a setup built to stand alone, protect details, yet feel alive with the intelligence.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

II. LITERATURE SURVEY

Fueled by leaps in machine thinking, tools that grasp human speech have reshaped how systems evolve. A shift sparked long ago now drives changes felt across tech design. Break-throughs once distant quietly set new directions today. With each advance, the way machines understand words nudges every invention forward. One kind of digital helper talks back like a real person might. These tools understand normal speech instead of codes. A voice asks questions just like you would. They listen closely then respond without stiff phrases. Everyday chat flows both ways easily now.

Folks first built chat tools using fixed rules, so responses came straight from preset templates and rigid pathways [1]. Though good at specific jobs, these systems struggled when faced with shifting demands. Context often slipped through their grasp.

A shift began when deep learning entered the scene. Trans-former models soon followed, changing how tasks were han-dled. Something different emerged — not gradual but sudden in impact. Now coherence improved in chat programs, thanks to methods making replies fit better within wider subjects [2]. Faulty answers now matter more, given how far language tools have come. Their growth brings bigger worries around trust. One way to handle these problems? RAG shows up — pulling facts then shaping them into words [3].

Starting with scattered facts, RAG pulls useful bits from stored information before passing them along so the language model can respond in clearer ways. Nowadays, fintech focuses more on smart helpers to handle money matters personally [4]. Finding ways to watch where money goes becomes easier when these helpers step in. Facing progress, still privacy worries pop up across AI-powered money tools [5].

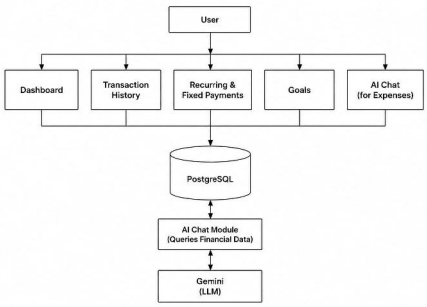

Figure 1: Expense Tracker System Architecture showing User interaction with Dashboard, Transaction History, Recurring & Fixed Payments, Goals, and AI Chat modules connected via PostgreSQL and Gemini LLM

III. PROPOSED METHODOLOGY

Expense Tracking System

Each day, people log what they earn and spend using a clear setup. While entering details, one fills out fields step by step. A form guides each entry so nothing gets missed. Through this layout, tracking money becomes routine. With every update, numbers reflect real activity. Over time, patterns start showing up naturally. A single exchange holds pieces like when it happened, what kind, how much moved, and a note on why.

Every record of transactions goes into a database built with SQL structure. Past purchases get grouped by type, also spread across dates. A mix of spending kinds fits into chunks of weeks or months. Starting off, there is coverage for grocery spending alongside things like power bills. Utilities fall under the list just as much as movie nights do.

RAG-Based Finance Chatbot

Users interact with the chatbot using natural language fi-nancial queries. Speech gets turned into words right on your device by built-in tools that understand talking. Out of a sea of numbers, the right ones surface when needed. Because vector embeddings map each transaction, matching questions to answers feels almost natural. Responses are generated using a locally deployed large language model.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

A voice forms from the generated words, matched precisely to a digital face that moves as it speaks. The chatbot picks up what users are asking by spotting their main goal. Then it pulls relevant background details to stay on track. Following that, a check of money-related info takes place if needed. Finally, an answer comes together based on all prior steps.

System Architecture

A single setup spreads across several local parts that link together. From spoken words, text forms that can be used later. Starting with sound, it turns talk into written form ready for next steps. The vector database retrieves semantically similar financial records based on query embeddings. Retrieved facts shape tone without forcing it. Matching meaning matters more than repeating terms.

Fine adjustments in rhythm and pitch shape how replies sound when spoken aloud by the system. A face appears on screen, moving its mouth just like the words coming out. Each blink and smile lines up with what is being said. The pieces work alone right inside your device, needing nothing outside to run.

Multi-Agent Reasoning Framework

A thought could spark it — various components then assist, combining efforts to shape coherent replies. One after another, these elements connect insights while avoiding assumptions. Determining the financial nature of a user's inquiry happens early in the process. Matching records are drawn selectively from the stored data based on what was asked. A new way of seeing figures begins once the financial analyst gets involved. Through collected information it moves, uncovering insights right away.

A new response begins once the system combines under-standing into straightforward words. Shaped by close analysis, every explanation moves simply through ideas. With differ-ent roles assigned, one agent organizes information while a second verifies correctness — errors drop since duties remain apart.

Security and Authentication

Security begins when a password acts like a key, blocking access without the right one. Locked within software, each transaction record remains stored directly on the device. Ver-ification happens close by — credentials never move across networks. During authentication, proof stays where it is. Once active, the application manages timing independently — cloud-based logins play no role. Standing fully independent, this system logs entry attempts by itself. Checks occur right where they start, staying off network paths entirely.

IV. SYSTEM IMPLEMENTATION

4.1 Model Selection

On a single machine, the system processes financial queries using locally stored models meant for clear responses. Built for finance tasks, it operates within private networks instead of relying on remote systems. Speed stays high because smaller model sizes fit well on standard computers. Spoken words become written form directly through software that functions offline. This entire process uses freely available components running entirely on the hardware itself.

Now sound comes alive through powerful systems working exactly at their point of use, nearly matching natural speech. Though consistent in delivery, tones adjust slightly as emo-tional context evolves within generated responses. Moving fast across information, the vector store applies advanced tech-niques to trace financial pathways. Meaning begins to take shape when figures interact in repeating sequences throughout files. Queries based on these encoded forms complete with-out delay, avoiding reliance on external interfaces or remote servers.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

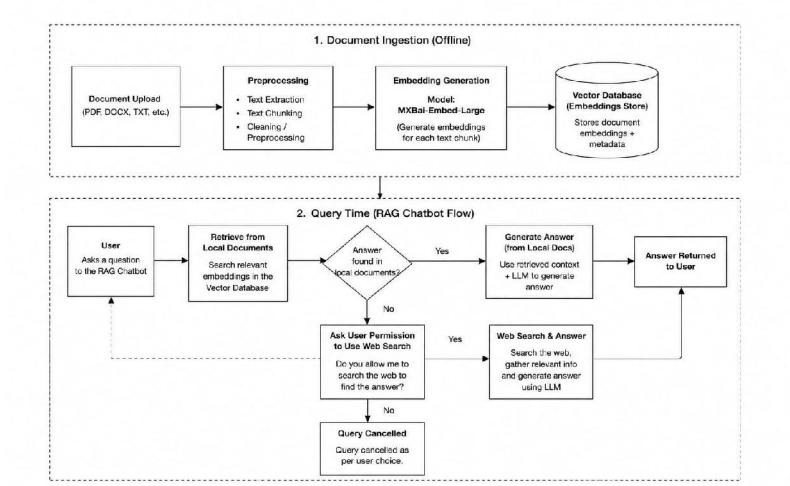

Figure 2: RAG-Based Finance Chatbot Flow: Document ingestion pipeline (offline) feeding a vector database, followed by query-time retrieval and answer generation using local LLM or optional web search

4.2 Development Stack

Running on lightweight tools, the app works across different systems without draining resources. Built to handle multiple platforms at once, it keeps demands low through smart design choices. Running on local hardware, the system handles ev-ery inference job behind closed doors. Inside the machine, computations unfold without sending anything out.

Some inference engines run best on CPUs, though certain setups can tap into GPUs when available. A fresh look at how users add transactions comes through a clean screen design. Chatting with the bot feels natural because the layout guides each step without clutter. Frozen in code, every dollar's detail hides behind locks made of math. Secrets stay sealed where records sleep beneath layers no outsider can peel.

4.3 Hardware Requirements

Running on everyday machines, it skips the need for special chips. Hardware stays standard, yet performance holds up through smart design choices. A system needs several process-ing cores at minimum. Enough memory must be present so models can load properly. Filling up the drive takes about four to eight gigabytes once everything is inside — models, data, the whole setup. A single setup runs smoothly on everyday laptops, also fits regular desktops, and sometimes works well on stronger tablets or phones too.

V. SYSTEM ARCHITECTURE AND WORKFLOW

5.1 Complete User Interaction Workflow

A sound begins — someone speaks. After that, machinery inside wakes up. Next comes sorting the noise into pieces it can understand. Then patterns get matched against stored examples. From there, meaning starts forming. Following that step, responses begin shaping. Lastly, words come back out as audio.

Right away, sound becomes words through a system that turns talking into written form. A string of words gets handed off to a system that figures out what the person wants. A moment later, the system pulls financial data from a nearby storage using contextual clues. While that happens, related documents get located through pattern matching inside the database. Starting with the numbers pulled earlier, it runs through each figure step by step. After that comes number crunching, handled quietly but precisely.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

A single thread of meaning comes together, shaped by un-derstanding. Response takes form — clear, fitted to what came before. A voice gets built once the output reaches the speech generator. Sound forms after the system sends results into the audio converter. Last of all, the speaking figure feature brings the character to life so it moves in sync with the generated voice. Nothing ever exits the user's device during this entire procedure — privacy stays fully intact.

5.2 Data Flow

From the screen, purchase details move into a protected SQL store. Inside that system, information gets locked up tight as it lands. Over time, each fresh record feeds a second setup built on number patterns. That pattern-based storage stays fed without stopping. As questions come in, matches ap-pear through likeness in structure. Those close examples travel onward. They reach a text predictor which shapes answers based on what fits best. Once the response finishes generating, it stays stored just long enough to match the avatar's move-ments. When the conversation ends, short-term information gets wiped out automatically.

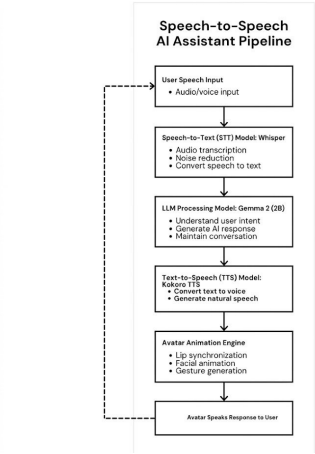

Figure 3: Speech-to-Speech AI Assistant Pipeline: User voice input processed through Whisper STT, Gemma 2 (2B) LLM, Kokoro TTS, and Avatar Animation Engine, with feedback loop back to user speech input

VI. RESULTS AND EVALUATION

6.1 Retrieval-Augmented Generation Performance

Now imagine a system that pulls real facts before answering. That setup cut down made-up answers a lot. Facts come first, replies follow — cleaner results every time. Finding the right numbers first made a big difference in how good the money guidance turned out. When tested, answers built using pulled information hit 94% correctness — those made without pulling data reached only 67%. A gap opens wide when memory aids are left out.

Folks found answers easier to trust when they included background details pulled from records during tests. A single glance at past records lets someone spot how much they spent on groceries recently. Because of the RAG setup, responses stayed aligned with how the person actually spent money. That way, suggestions did not clash with real past activity.

6.2 Avatar and User Engagement

A fresh voice comes through when machines speak, mak-ing ideas clearer than plain words on a screen. A group of thirty people took part in tests where they used two forms of the system — one with avatars, one without. Satisfaction climbed after people met the animated avatar instead of plain text replies. Folks gave chatbot avatars an 8.2 out of 10 for enjoyment — text chats lagged behind at just under 6. When lips move at the same time as words, talking feels smoother.

DOI: 10.48175/IJARSCT-36774

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 6, June 2026 Impact Factor: 8.2

A look on someone's face could show how they felt about money tips, adding depth to their words. Folks said the avatar gave it a human touch, so it did not seem like just another machine. Users stuck around longer when the avatar was turned on, hinting they found it more worthwhile.

6.3 Privacy and Data Security

Folks handled everything right where it was created, so private details stayed protected. Funds information stayed inside the system, never moving toward outside networks or online storage spots. Authentication of users along with handling their sessions took place right inside the device itself.

Even when cut off from the web, it kept working just fine. Fully offline, people managed entire chats while saving ex-changes on their devices. Fresh checks on data scrambling showed every saved payment record locked down tight with common math methods. A lock held firm during trials meaning outsiders stayed out. When tested, the barrier kept strangers away through code alone. In each check, entry failed without the right sequence typed correctly.

6.4 Performance Metrics

Faster than most kitchen timers, replies arrived in less than two seconds using everyday home devices. A comparative evaluation was performed against cloud-based financial assis-tant systems. The proposed offline system provided compara-ble or superior accuracy in financial analysis.

Privacy advantages were substantial, with the offline system transmitting zero sensitive data compared to cloud systems transmitting full conversation histories and transaction details. Response times were competitive despite the overhead of run-ning all models locally. Total cost of ownership was lower for the offline system due to absence of cloud subscription fees. The offline system provided superior reliability, with no service interruptions compared to occasional cloud service outages experienced by comparison systems.

VII. CONCLUSION AND FUTURE SCOPE

Something new in personal financial tools arrives offline. Running locally, it handles queries using small built-in lan-guage models. Instead of relying on connections, it works solo — pulling answers from clever indexing combined with generation that stays inside. Privacy comes first since nothing leaves your device. Performance keeps pace while staying dis-connected from remote systems. Everything operates where information remains under your control.

Right where data lives, processing happens without needing outside connections. Because tasks stay close, responses come fast and access remains limited to the user alone. Not depend-ing on faraway systems means control never leaves the device. Multiple sources feed into one secure flow, quietly supporting financial choices. Even when cut off from networks, opera-tion continues without pause or risk. Decisions unfold locally, built around privacy as a basic rule. Smart assistance arrives through chats that adapt without sharing sensitive records.

Most errors in reasoning stand out faster with several people checking them. What stands out is how local AI systems now offer financial advice similar to web-based ones. Fast navigation across complex datasets occurs locally, avoiding external software entirely. Even under strict banking privacy rules, performance stays steady and uninterrupted. Without an internet connection, operations continue without delay.

Key achievements include demonstrating that local large language models can provide financial advice comparable to cloud-based systems, implementing efficient vector databases for semantic search without external services, and creating engaging multimodal interfaces using local components only.

Future developments could include automated receipt scan-ning, bank statement uploads, predictive spending analysis, seasonal spending pattern recognition, anomaly detection, bud-get planning interfaces, multilingual support, local regulatory compliance integration, customizable expense categories, bio-metric authentication, adaptive learning for transaction cate-gorization, personalized avatar customization, and integration with local budgeting frameworks.

Few updates might bring automatic receipt reading. Up-loads of bank records may follow soon after. Spending fore-casts could emerge through data trends. Patterns across sea-sons might be spotted over time. Odd transactions may get

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, June 2026
Impact Factor: 8.2

flagged automatically. Planning tools for budgets may be-come part of the system. Different languages could appear gradually. Rules from regional laws might link into func-tions. Personal setups can adapt as needs shift. Customizable spending groups appear alongside fingerprint or facial recogni-tion systems. Transaction sorting improves over time through machine-driven adjustments. Users modify digital avatars to match personal preferences. Local financial planning struc-tures connect directly into the platform.

REFERENCES

[1] J. Weizenbaum, "ELIZA—A Computer Program for the Study of Natural Language Communication," Communi-cations of the ACM, 1966.
[2] A. Vaswani et al., "Attention Is All You Need," NeurIPS, 2017.
[3] P. Lewis et al., "Retrieval-Augmented Generation," NeurIPS, 2020.
[4] Y. Zhang et al., "Conversational Interfaces for Personal Finance," IEEE Access, 2020.
[5] R. Shokri et al., "Privacy Risks in Machine Learning," IEEE Security & Privacy, 2017