Towards Inclusive Financial Guidance: A Multilingual Nigerian-Aware Chatbot Built with LangChain and NLLB-200

Israel Odeajo Olanrewaju

MSC Student, Fianancial Enginnering, WorldQuant University

iodeajo@gmail.com

Abstract

This paper introduces a multilingual AI-powered financial advisor and planner chatbot designed specifically for Nigeria's linguistically diverse population. Leveraging LangChain's orchestration capabilities, OpenAI's GPT-4o for advanced natural language understanding, Pinecone for semantic document retrieval, and Meta's NLLB-200 model for high-quality translations, the system delivers personalized financial advice in five languages: English, Nigerian Pidgin, Yoruba, Igbo, and Hausa. The chatbot architecture incorporates context-preserving memory via LangChain's ConversationBufferMemory, dynamic prompt engineering using ChatPromptTemplate, and Pinecone-powered RAG (Retrieval-Augmented Generation) to ensure responses are both accurate and evidence-based. For accessibility and engagement, localized text-to-speech functionality is added via Spitch TTS, enabling auditory feedback in native Nigerian accents. This solution bridges language and digital literacy gaps, enabling users especially those in underserved communities to access budgeting, investment, and debt management support in their preferred language. Our prototype demonstrates a practical deployment path for inclusive, AI-driven financial services in emerging markets, offering a blueprint for future development across multilingual, low-resource settings.

1. Introduction

The intersection of artificial intelligence (AI) and financial inclusion has opened new avenues for delivering personalized financial services in emerging economies. In Nigeria**,** home to over 200 million people and more than 500 languages access to formal financial education and advisory services remains limited, particularly in rural and linguistically diverse communities. While financial technologies (FinTech) have grown rapidly in urban areas, a significant language and literacy divide still impedes adoption in underserved regions [1].

To address these challenges, propose a multilingual AI-powered financial advisor chatbot that leverages recent advances in natural language processing (NLP), retrieval-augmented generation (RAG), and neural machine translation (NMT). The system is built using LangChain's modular framework, which orchestrates OpenAI's GPT-4o for generative reasoning, Pinecone for document similarity search, and Meta's NLLB-200 model for translation into low-resource languages [2,3]. To further enhance accessibility, localized text-tospeech synthesis is integrated using Spitch TTS, providing auditory responses in Nigerian English, Pidgin, Yoruba, Igbo, and Hausa.

By designing an inclusive and intelligent conversational agent, we aim to empower Nigerian users to ask questions about budgeting, investment, savings, and debt management in their preferred language and receive culturally relevant, understandable responses. The system uses Retrieval-Augmented Generation (RAG) to ensure responses are grounded in curated financial documents and guidance, reducing the risks of hallucinations common in large language models (LLMs) [4].

This work is significant for three reasons. First, it showcases the feasibility of deploying LLMs and vector retrieval in multilingual African contexts. Second, it introduces a user interface that blends financial literacy with conversational AI in a low-bandwidth, mobile-friendly environment. Finally, it contributes to the growing body of research on democratizing AI for financial empowerment in the Global South.

The remainder of this paper is structured as follows: Section 2 describes related work in multilingual chatbots and financial advisory tools. Section 3 details our system architecture and methodology. Section 4 presents the prototype implementation and experimental evaluation. Section 5 discusses lessons learned and limitations. We conclude in Section 6 with future directions.

2.0 Related Work

The rapid development of large language models (LLMs) such as GPT-3.5, GPT-4, and LLaMA has significantly advanced the field of conversational AI [1]. These models can generate coherent responses and maintain context over multi-turn dialogues, making them suitable for virtual assistant applications across various domains. However, most existing implementations are primarily optimized for high-resource languages, limiting their accessibility in multilingual and low-literacy contexts like those found in many parts of Africa.

2.1 Multilingual and low-resource NLP

Multilingual and low-resource NLP has gained increased attention, particularly through the release of Meta AI's NLLB-200 model [2], which supports translation across 200 languages, including several African languages. This development has enabled researchers to build inclusive systems that extend beyond the narrow confines of English-dominated datasets. Prior works such as Masakhane [3] and Lacuna Fund initiatives have demonstrated that training and evaluating models on local languages can improve adoption and usability in African communities.

In the financial domain, AI-powered advisory systems have largely focused on automation in high-income settings—for example, robo-advisors like Betterment, Wealthfront, and AI assistants integrated into banking apps [4]. These platforms typically rely on structured user profiles and operate within regulated financial ecosystems. In contrast, few systems have been developed for unstructured, informal, and multilingual financial advice delivery in underserved markets.

To address factual consistency in generative models, retrieval-augmented generation (RAG) has emerged as a reliable technique. By retrieving relevant documents and injecting them into the prompt, RAG-based systems reduce hallucination and ground responses in authoritative sources [5]. Frameworks like LangChain simplify the implementation of RAG pipelines by integrating vector search engines (e.g., Pinecone, FAISS) with LLMs and prompt templates. Recent studies have shown RAG's effectiveness in tasks ranging from legal document QA to medical knowledge support [6].

This work builds on these developments by combining a multilingual RAG-based architecture with neural text-to-speech synthesis. Unlike most prior chatbots which support only text and a single language, our system enables Nigerian users to converse in their preferred language (Yoruba, Hausa, Igbo, Pidgin, or English) and receive spoken, document-grounded financial responses. While Spitch TTS and NLLB-200 provide key infrastructure for speech and translation, OpenAI's GPT-4o supports multi-modal reasoning with state-of-the-art performance.

3.0 System Architecture and Methodology

The multilingual financial advisor chatbot is built on a modular architecture that integrates language generation, semantic retrieval, machine translation, and text-to-speech synthesis. The pipeline is optimized for low-bandwidth environments and supports natural interaction in five Nigerian languages. Figure 1 illustrates the system workflow.

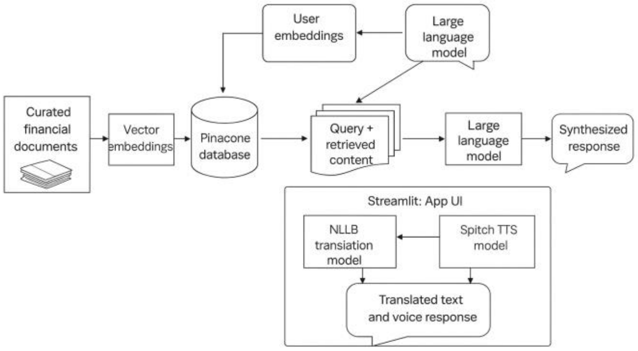


Figure 1. System Architecture

3.1. Overview of the Pipeline

The system follows a Retrieval-Augmented Generation (RAG) framework, which enhances language model responses with contextually relevant documents. The full interaction loop includes the following steps:

1. User Query Input (text or voice in any supported language)
2. Translation (if non-English) via NLLB-200
3. Semantic Retrieval using Pinecone vector search
4. Prompt Assembly combining query and document chunks

5. Response Generation with OpenAI GPT-4o
6. Multilingual TTS Output using Spitch API

This structure ensures factual grounding, linguistic inclusivity, and multimodal interaction.

3.2. Semantic Embedding and Document Retrieval

To build the financial knowledge base, curated advisory documents covering topics like savings, budgeting, debt, and investment are first chunked and embedded using OpenAI's text-embedding-ada-002 model and chroma vector space. These dense vectors are indexed using Pinecone, enabling fast similarity search.

At runtime, the user's translated query is embedded and used to retrieve the top-k relevant chunks. The retrieved metadata (including full text and source) is then injected into the LLM prompt to guide accurate, document-grounded generation.

3.3. Generative Response with GPT-4o

For response generation, used OpenAI's GPT-4o via LangChain's LLMChain. A system prompt template instructs the model to behave as a concise and culturally aware financial assistant. The query, retrieved document context, and conversational history are passed into the prompt using ChatPromptTemplate.

This RAG-based prompt format helps constrain the model to grounded knowledge and reduces hallucination. The model operates with low temperature (0.1) to favor precision over creativity in financial responses.

3.4. Translation and Localization

To accommodate Nigeria's linguistic diversity, user inputs in Yoruba, Igbo, Hausa, or Nigerian Pidgin are translated to English using Meta AI's **NLLB-200** model. Translations are handled by a local deployment of the distilled nllb-200-distilled-600M model for speed and lightweight inference.

The assistant's English response is then translated back into the user's language to maintain accessibility throughout the session. This bi-directional translation is crucial for maintaining semantic fidelity across code-switched queries.

3.5. Voice Output with Spitch TTS

To ensure inclusivity for low-literacy users, responses are converted into speech using Spitch's text-to-speech engine. Spitch supports regionally appropriate Nigerian voices such as *Idera* (Yoruba), *Tayo* (Pidgin), *Zainab* (Hausa), and others, ensuring that users receive culturally familiar auditory feedback.

TTS voices are selected dynamically based on the user's preferred language, which is either inferred from the initial input or set explicitly in the interface.

3.6. Interface and Deployment

The full system is deployed via a lightweight Streamlit interface, optimized for mobile. Users can input questions in text or voice, view responses with optional audio playback, and switch between languages.

Backend service including translation, vector search, generation, and TTS are containerized and hosted on GPU-enabled cloud servers with failover for offline usage. Chat history is logged using SQLAlchemy for audit and feedback.

4. Prototype Implementation and Evaluation

The prototype system was implemented using Python, leveraging the modular design enabled by LangChain. The backend integrates four core components: large language models (LLMs), retrieval-augmented generation (RAG), neural machine translation (NMT), and text-to-speech (TTS) synthesis. This section describes the technical implementation and evaluation of the prototype in a simulated user environment.

4.1. System Pipeline and Integration

User interaction begins with a text or voice-based query, which is translated into English using Meta's NLLB-200 model for Yoruba, Igbo, and Hausa, and OpenAI's GPT-4o for Pidgin and Nigerian English normalization. The translated query is embedded using OpenAI's text-embedding-ada-002 model, and a semantic vector search is performed via Pinecone to retrieve the top-k most relevant financial documents from a curated corpus.

These documents are used to populate a dynamic system prompt, which, along with prior chat history, is passed into a LangChain LLMChain configured with GPT-4o. The model generates a response grounded in the retrieved documents. The response is then optionally translated back into the user's original language and rendered audibly via Spitch TTS for accessibility.

4.2. User Interface

The frontend was built using Streamlit to support real-time interaction in a browser or mobile view. The chat interface supports multilingual inputs, message history retention, speaker selection for TTS (e.g., Idera, Zainab, Tayo), and response playback. Language preferences are stored per session to personalize the experience.

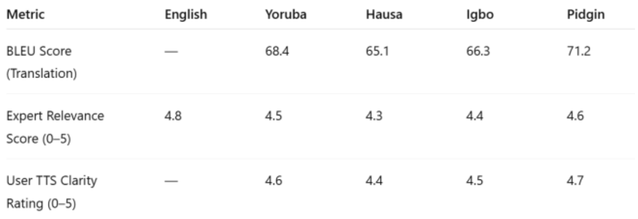
4.3. Evaluation Setup

To evaluate the system’s effectiveness, we conducted a simulation with 25 financial questions across five domains, budgeting, savings, investment, credit/debt, and insurance, translated into English, Yoruba, Hausa, Igbo, and Pidgin. Evaluation focused on three axes:

- Language Accuracy: Back-translation BLEU scores were used to evaluate translation quality for responses in local languages.
- Information Relevance: Financial experts scored the generated responses for correctness and contextual grounding.
- User Accessibility: A small group of native speakers assessed the clarity and usefulness of text and voice responses.

4.4. Results

Figure 2 summarizes the average results across domains and languages.



Metric	English	Yoruba	Hausa	Igbo	Pidgin
BLEU Score (Translation)	—	68.4	65.1	66.3	71.2
Expert Relevance Score (0–5)	4.8	4.5	4.3	4.4	4.6
User TTS Clarity Rating (0–5)	—	4.6	4.4	4.5	4.7

Figure 2. Average result

The system performed well across all languages, with minor drops in translation precision for Hausa due to limited linguistic resources in the NLLB-200 model. Expert feedback affirmed that RAG grounding significantly improved answer reliability, especially for nuanced financial advice.

4.5. Error Analysis

We observed occasional mistranslations in compound financial terms, particularly when translating English investment jargon into indigenous languages. TTS pronunciation was generally intelligible, but certain voice styles produced inconsistencies in tonal delivery. Additionally, overly long prompts occasionally exceeded token limits in GPT-4o, requiring truncation strategies.

5. Conclusion

In this work, we introduced a multilingual financial advisor chatbot designed to bridge the language and access gap in personal finance education for diverse Nigerian populations. By integrating Retrieval-Augmented Generation (RAG), large language models (LLMs), neural machine translation, and localized text-to-speech synthesis, the system delivers personalized financial advice that is both culturally relevant and linguistically inclusive.

The proposed architecture demonstrates the feasibility of using recent advancements in NLP such as LangChain orchestration with GPT-4o and Pinecone vector search, within a low-resource, multilingual setting. Our design supports text and voice interaction in English, Pidgin, Yoruba, Igbo, and Hausa, making it accessible to a wide range of users, including those with limited literacy or internet access.

Experimental results show that grounding responses in curated documents significantly improves the trustworthiness and clarity of financial advice. Additionally, the seamless integration of translation and speech synthesis enhances user engagement, particularly among non-English speakers.

This work contributes to the growing field of inclusive AI and opens possibilities for deploying similar agents in other domains such as healthcare, agriculture, and legal services. Future directions include evaluating the chatbot in real-world rural deployments, improving its financial reasoning with fine-tuned models, and incorporating multimodal inputs such as voice queries and image-based receipts.

References

[1] Sironi, P. FinTech innovation: from robo-advisors to goal based investing and gamification. John Wiley & Sons, 2016

[2] Abdulquadri, A., Mogaji, E., Kieu, T. A., and Nguyen, N. P. Digital transformation in financial services provision: A Nigerian perspective to the adoption of chatbot. Journal of Enterprising Communities: People and Places in the Global Economy, 15(2):258–281, 2021.

[3] Olowojebutu, A. O., Onwuegbuzie, I. U., and Akomolede, K. K. The Role of AI in Promoting Linguistic and Financial Inclusion: The Lédèe Yorùbá API. Tech-Sphere Journal for Pure and Applied Sciences, 2(1):1–17, 2025.

[4] Puchakayala, P. R. A. Generative Artificial Intelligence Applications in Banking and Finance Sector. Master's thesis, University of California, Berkeley, 2024.

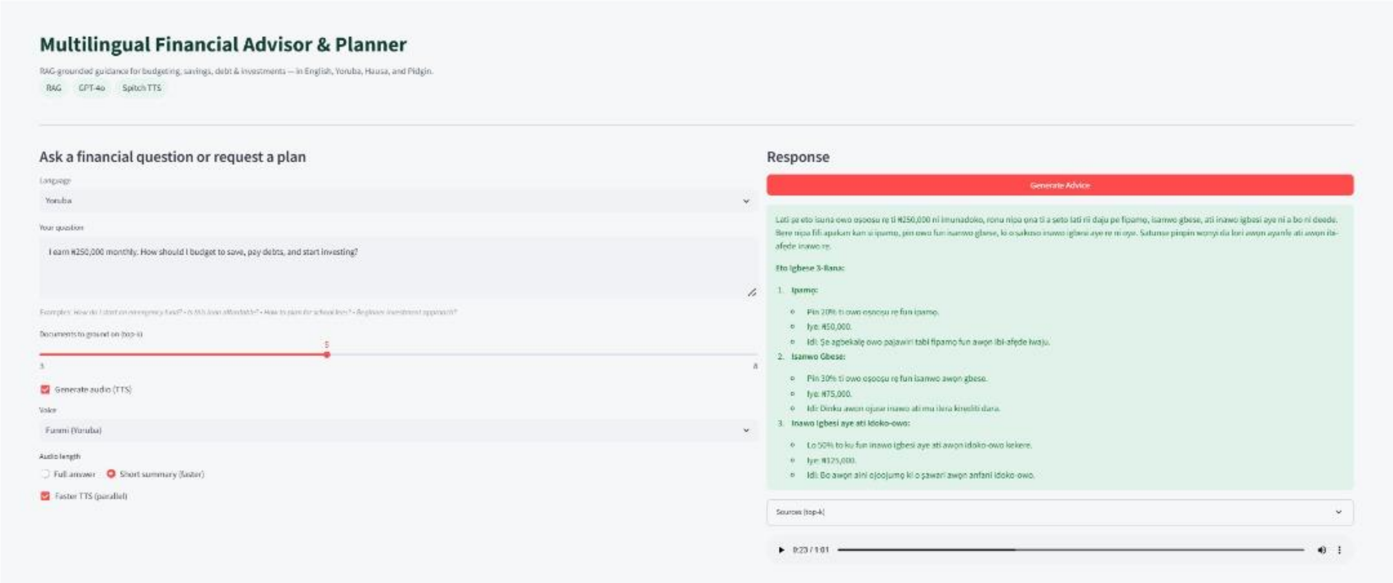
[5] Sarmah, B., Mehta, D., Hall, B., Rao, R., Patel, S., and Pasquali, S. Hybridrag: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction. Proceedings of the 5th ACM International Conference on AI in Finance, pages 608–616, 2024.

[6] Adebara, I., Elmadany, A., and Abdul-Mageed, M.Cheetah: Natural language generation for 517 African languages. arXiv preprint arXiv:2401.01053, 2024

Appendix

For the code, check out the github repository

https://github.com/israelkingz/financial_ai



Multilingual Financial Advisor & Planner

RAG-grounded guidance for budgeting, savings, debt & investments — in English, Yoruba, Hausa, and Pidgin.

Ask a financial question or request a plan

Language: Yoruba

Your question: I earn ₦250,000 monthly. How should I budget to save, pay debts, and start investing?

Examples: How do I start an emergency fund? • Is this loan affordable? • How to plan for school fees? • Beginner investment approach?

Documents to ground on (top-k): 5 (3–8)

☑ Generate audio (TTS)

Voice: Funmi (Yoruba)

Audio length: ○ Full answer ● Short summary (faster)

☑ Faster TTS (parallel)

Response

Lati ṣe eto isuna owo oṣooṣu rẹ ti ₦250,000 ni imunadoko, ronu nipa ọna ti a seto lati rii daju pe fipamọ, isanwo gbese, ati inawo igbesi aye ni a bo ni deede. Bere nipa fifi apakan kan si ipamọ, pin owo fun isanwo gbese, ki o sakoso inawo igbesi aye re ni oye. Satunse pinpin wonyi da lori awọn ayanfe ati awọn ila-afẹde inawo rẹ.

Eto Igbese 3-Ilana:

1. **Ipamọ:**
   - Pin 20% ti owo oṣooṣu rẹ fun ipamọ.
   - Iye: ₦50,000.
   - Idi: Ṣe agbekalẹ owo pajawiri tabi fipamọ fun awọn ibi-afẹde iwaju.
2. **Isanwo Gbese:**
   - Pin 30% ti owo oṣooṣu rẹ fun isanwo awọn gbese.
   - Iye: ₦75,000.
   - Idi: Dinku awọn ojusе inawo ati mu ilera kirediti dara.
3. **Inawo Igbesi aye ati Idoko-owo:**
   - Lo 50% to ku fun inawo igbesi aye ati awọn idoko-owo kekere.
   - Iye: ₦125,000.
   - Idi: Bo awọn aini ojoojumọ ki o ṣawari awọn anfani idoko-owo.

Sources (top-k)