**Hong Kong Institute of Vocational Education**

**Discipline of Information Technology**

**IT114116 – HD in Data Science and Analytics**

**Final Year Project – Big Data Analytics (ITP4870M)**

**Project Proposal (AY2025/26)**

**FraudSim: Immersive Fraud Awareness Through RAG-Guided LLM Simulation**

**Group Members:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Name (Student ID)** | **E-Mail** | **Phone Number** |
| 1. | TAN James Anthroi (240350922) | 240350922@stu.vtc.edu.hk | 5911-5923 |
| 2. | AQSA-AHMED (240044574) | 240044574@stu.vtc.edu.hk | 6433 4130 |
| 3. | LIN Yueying (240444846) | 240444846@stu.vtc.edu.hk | 5958-4735 |
| 4. | TAN Xiuhao (240253372) | 240253372@stu.vtc.edu.hk | 6768-6457 |

# **Introduction to the proposed project (Statement of problem to be solved)**

Fraud remains a widespread and evolving threat, impacting individuals and institutions globally. In Hong Kong and elsewhere, vulnerable groups such as the elderly and youth are frequently targeted due to limited awareness or digital literacy. Traditional fraud prevention methods—like posters, seminars, and announcements—often lack engagement and personalization. This project proposes an immersive AI-powered education system that simulates realistic fraud scenarios tailored to each user’s context, enhancing awareness through experiential learning. The system leverages factual data and real-world cases to ensure authenticity and relevance.

# **Background (Background of the problem)**

Despite ongoing awareness campaigns, fraud continues to cause significant financial and emotional damage. Victims often report being unaware of scam tactics or failing to recognize warning signs. Existing educational tools are static and impersonal, making it hard for users to relate. Language barriers and generational gaps further reduce effectiveness. This project addresses these issues by using real fraud data—scraped from credible sources—and AI technologies to create dynamic, personalized learning experiences. The system ensures that users receive accurate, grounded information in their preferred language.

# **Outline of proposed solution**

The proposed solution is a web-based platform that uses Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) to simulate personalized fraud scenarios. Users begin by recording a short self-introduction, which is transcribed using Whisper (speech-to-text with multilingual support). This input is vectorized via ChromaDB and securely stored. The system then uses RAG to retrieve relevant fraud cases from public datasets and web-scraped sources (e.g., Reddit, HK01, Yahoo, HKMA), guiding the LLM to generate realistic scam simulations. LangChain agents orchestrate the interaction, enabling structured prompts and function calling for deeper analysis.

**Core Features**

* Personalized Scam Simulation
* Multilingual Support
* Real Case Grounding via Web Scraping
* LangChain Agent Integration
* Optional Avatar Interaction
* Summary Report with Objective Metrics

**Algorithms**

The system will adopt a layered hybrid approach to simulate and analyze fraud scenarios effectively:

* **Named Entity Recognition (NER) + RAG + LLM (Primary):** Detect fraud-related entities (e.g., scam types, impersonation cues, financial terms) from user input or transcribed speech. RAG ensures that the LLM responses are grounded in verified fraud case data sourced from curated datasets and web-scraped platforms.
* **Rule-Based Scoring Algorithm (Primary):** Cross-references detected entities with a structured fraud case database and assigns risk levels or scenario types based on predefined criteria (e.g., scam frequency, victim profile).
* **LangChain Agent (Orchestration Layer):** Manages prompt flow, function calling, and contextual memory across interactions. This agent ensures that the LLM operates with structured logic and traceable data sources, enhancing reliability and transparency.
* **Convolutional Neural Network (CNN) (Secondary):** Used optionally for image-based fraud detection (e.g., identifying fake documents or scam-related visuals). A pre-trained model (e.g., MobileNet or EfficientNet) will be adapted for lightweight deployment if needed.

This hybrid algorithmic structure ensures both feasibility and extensibility, allowing future multimodal enhancements while maintaining factual accuracy and user relevance.

**Technologies**

* **Frontend:** React (web platform) with optional avatar interface
* **Backend:** Node.js (API), MongoDB (vector and user data storage)
* **AI Pipeline:**
  + **RAG + LLM** (via LangChain, Qwen, or LlamaIndex) for fraud scenario generation, multilingual support, and personalized simulation
  + **Embedding Models** (e.g., Sentence-BERT, OpenAI embeddings) for semantic search and user profile matching
  + **LangChain Agent** for orchestrating prompt logic, source grounding, and function calling
  + **CNN (TensorFlow Lite)** for optional image-based fraud detection (secondary feature)
* **Speech & Text Processing:** Whisper for multilingual speech-to-text conversion
* **OCR (if needed):** Google ML Kit for extracting text from images or documents
* **CI/CD & Deployment:** Docker for containerization, GitHub Actions for automated testing and deployment

This tech stack ensures scalability, multilingual accessibility, and privacy compliance, while enabling real-time, personalized fraud education experiences.

# **Explanation of why proposed solution is appropriate**

This solution overcomes the limitations of traditional fraud education by offering interactive, personalized simulations grounded in real data. By integrating RAG and LangChain agents, the system ensures that LLM outputs are based on factual, sourced information rather than generic or fabricated content. Multilingual support and web-based deployment make the platform accessible and scalable. Objective reasoning—such as fraud frequency statistics and financial impact analysis—adds credibility to the learning experience, making it both informative and memorable.

# **Main development phases and Main Deliverables (Main Stages)**

|  |  |  |  |
| --- | --- | --- | --- |
| Phase | Timeline | Task | Assessment Element |
| 1 | Sep 2025 | Define scope, assign roles, finalize RAG+LLM+LangChain stack | Proposal (Due: 19 Sep 2025) |
| 2 | Oct–Nov 2025 | Literature review, collect fraud datasets, design user interface | Initial Report (Due: 28 Nov 2025) |
| 3 | Dec 2025 – Jan 2026 | Build RAG pipeline, integrate LLM and LangChain agents | Interim Report (Due: 9 Feb 2026) |
| 4 | Jan–Feb 2026 | Deliver working prototype with personalized fraud simulation | Interim Prototype & Demo (Due: 9–13 Feb 2026) |
| 5 | Mar–Apr 2026 | Add avatar interaction, conduct user testing | Final Report (Due: 18 May 2026) |
| 6 | May 2026 | Present fully functional platform with summary report generation | Final Prototype & Demo (Due: 18–22 May 2026) |
| Ongoing Performance & Teamwork | Throughout | Continuous team collaboration and performance tracking | Student Performance 10% |

# **Main deliverables**

* Initial, Interim, and Final Reports
* Executable Web-Based Platform
* Source Code with Documentation
* REST API and Backend Database
* Demo Presentation and User Testing Summary
* Fraud Data Analysis with Source References

# **The responsibilities of each member**

* **TAN James Anthroi A. (Team Leader & Backend Developer):** Oversees project coordination, backend API development, and secure data flow.
* **Mobile Developer:** Builds the React Native app, implements the chat interface, OCR integration, and ensures smooth user experience across Android/iOS.
* **Machine Learning Engineer:** Develops the RAG+LLM pipeline for allergen detection, trains CNN models for local cuisine recognition, and integrates AI into the app.
* **TAN Xiuhao (Backend Developer):** **(Backend Developer):** Designs server-side architecture, manages vector database, and supports real-time interaction.
* **LIN Yueying (QA & Testing):** Creates test plans, conducts user trials, documents bugs, and ensures system reliability.
* **Documentation & Reporting Specialist:** Prepares all written deliverables, maintains records, and supports presentations.

# **Executive Summary**

This project proposes an immersive anti-fraud education platform that simulates realistic scam scenarios using RAG+LLM and LangChain agents. Users begin with a short self-introduction, which is used to generate personalized fraud simulations grounded in real-world cases sourced via web scraping. The goal is to raise awareness and educate citizens—especially vulnerable groups—through interactive, data-driven learning. Technologies include Whisper, ChromaDB, LangChain, and optional avatar generation. The platform emphasizes scalability, multilingual support, and privacy, making it a practical and impactful tool for fraud prevention education.