HyperFin AI: Personalized Financial Intelligence for Nova Bank

Executive Summary

HyperFin AI is an intelligent, modular platform built to deliver truly personalized banking experiences — not just based on who a customer is, but on how they feel, what they do, and what they need. Developed for Nova Bank, the platform integrates machine learning models, large language models (LLMs), real-time behavior monitoring, and sentiment-aware systems to enhance customer engagement, product matching, and retention strategies.

Core Functionalities

1. Smart Product Recommendations

HyperFin AI analyses each customer's profile — including income, spending patterns, interests, and lifestyle — to recommend the most suitable banking products. Every recommendation is personalized, eligibility-checked, and backed by reasoning that makes sense to both users and advisors.

2. Retention-Driven Offer Engine

When a customer shows signs of dissatisfaction or churn risk, HyperFin AI doesn't wait. It actively evaluates their sentiment, past feedback, reviews, and tone. If frustration is detected, → A retention offer is instantly triggered (e.g., rate match, premium support). If satisfaction is detected, → A loyalty reward is proactively delivered (e.g., cashback, upgrades).

3. LLM-Powered Financial Q&A Assistant

This module identifies questions or confusion within customer comments on social media and responds with concise, trivia-style explanations that promote financial literacy.

Alongside each explanation, it recommends a relevant product tailored to the user's need, transforming feedback into personalized and educational engagement.

4. Fraud Detection with Explainability

Every transaction is evaluated not only for financial validity but also for behavioral consistency. This module detects anomalies in spending amount, payment methods, or merchant category, such as a sudden wire transfer to a luxury store after months of grocery transactions. If flagged, the system provides a detailed risk score, a human-readable explanation, and a recommended action (e.g., request verification).

5. Banking Trivia Module

This module analyses customer comments from social media to detect questions or confusion around banking concepts or services. When such a query is identified, the system responds with a short, informative trivia-style explanation related to the topic, enhancing financial literacy. Alongside the trivia, it recommends a relevant product that addresses the customer's concern, creating a personalized and engaging experience.

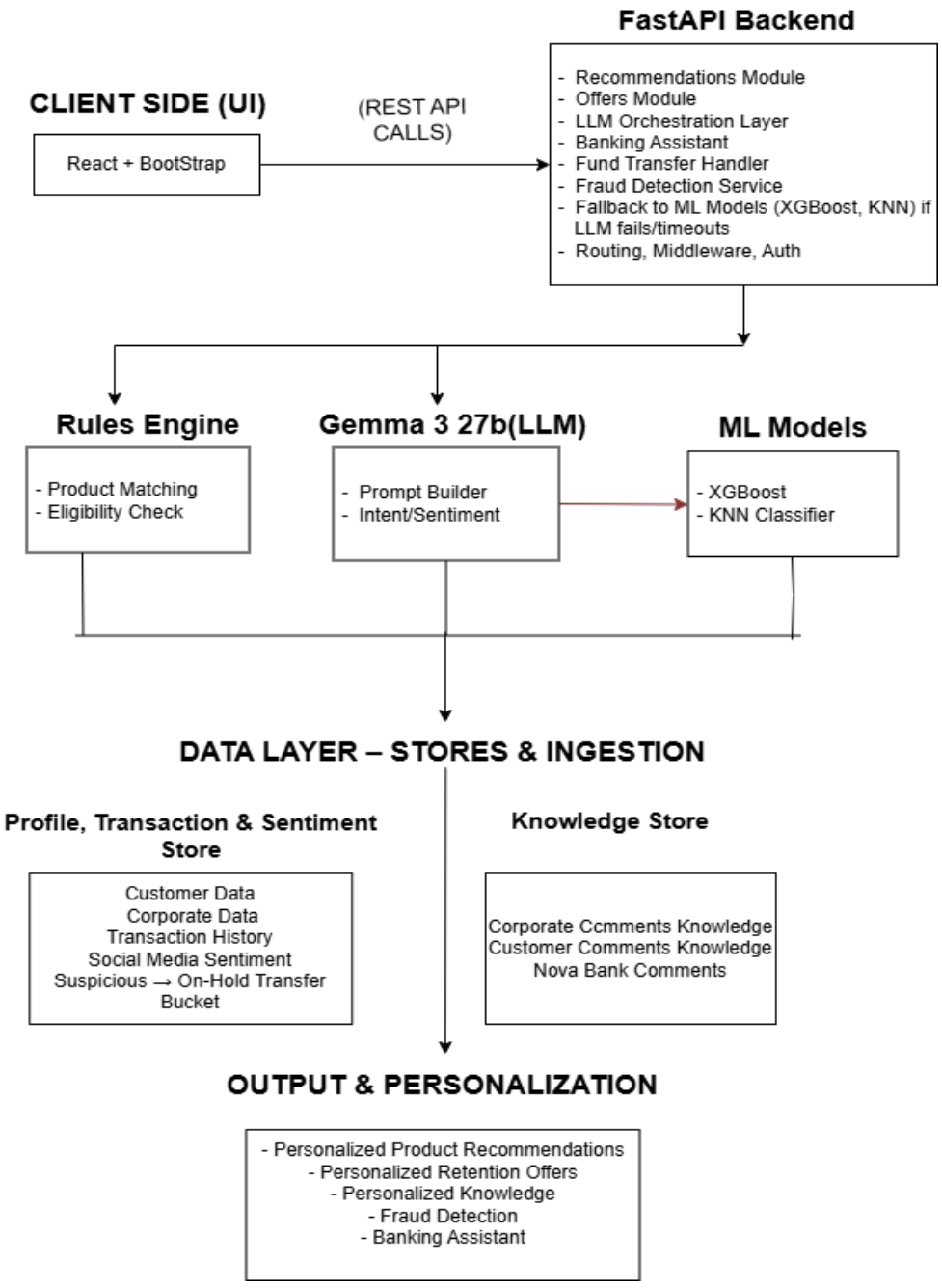
6. ML-Based Fallback Engine (XGBoost + KNN)

In scenarios where the LLM fails, times out, or lacks sufficient context, the system seamlessly falls back to robust ML models. XGBoost classifies churn risk using behavioural signals, while KNN identifies similar users for product recommendations. This ensures continuity, reliability, and intelligent outcomes — even in edge cases.

7. Real-Time Behavior Update via Fund Transfers

Every fund transfer initiated by a customer is logged instantly and used to update their transaction profile. This enables the system to dynamically adjust product recommendations, churn risk, and reward strategies based on new financial behavior. It ensures personalization is not static — it evolves in real-time with every user action.

System Architecture



Approach & Methodology

HyperFin AI was architected with an **LLM-first design philosophy**, where the large language model (Gemma 3 27B) serves as the core reasoning engine for understanding customer behaviour, generating product recommendations, interpreting sentiment, and answering natural language queries. Our goal was to enable Nova Bank to deliver human-like, adaptive personalization across all customer touch points. The system relies on the LLM to orchestrate decisions using customer profiles, transaction summaries, and emotional tone extracted from social feedback. To ensure robustness, we implemented a **fall-back strategy** using traditional ML models: XGBoost for churn prediction and KNN for collaborative filtering-based recommendations. These models are invoked when the LLM encounters edge cases, such as low-confidence outputs, ambiguous input, or timeout conditions.

Our modular pipeline includes:

- **LLM layer**: Executes primary tasks such as recommendation reasoning, churn explanation, financial Q&A, and behaviour-driven reward logic.
- **Sentiment processor**: Scores social media inputs for emotion-aware personalization.
- **Fall-back ML layer**: Ensures continuity of service by returning fast, reliable results when LLM logic is unavailable.
- **Real-time behaviour module**: Captures fund transfers and updates user state on the fly.

The development cycle followed a hybrid of data exploration, prompt engineering, and iterative evaluation. This allowed us to maintain high personalization accuracy, conversational intelligence, and real-time responsiveness—all while ensuring explainability and resilience in production-like settings.

Model Selection

HyperFin AI uses a **dual-layer architecture** combining a powerful **LLM (Gemma 3 27B)** as the **primary reasoning engine**, with traditional machine learning models (XGBoost and KNN) as **fallback mechanisms**.

- **Gemma 3 27B**: Chosen for its instruction-tuned architecture, domain adaptability, and robust handling of natural language tasks such as personalized financial Q&A, sentiment reasoning, and user profiling.
- **XGBoost**: Used for churn classification. Selected due to its gradient boosting capability, interpretability (via feature importance), and ability to handle missing/ tabular data efficiently.
- **KNN**: Applied in the recommendation engine for identifying similar customer profiles and generating personalized product suggestions when LLM confidence is low or unavailable.

This hybrid approach ensures the system remains accurate, explainable, and resilient — even in edge cases.

DATA LAYERS

HyperFin AI is fuelled by a rich, multi-dimensional dataset created and curated using LLMs to simulate real-world banking behaviour and diversity. We've designed this synthetic dataset to mirror realistic user profiles, transaction behaviour, feedback sentiment, and product logic — enabling powerful downstream AI decisions.

1. **Individual Customer Profiles**: Generated using LLMs to reflect variations in: Age, gender, income level, Credit score ranges, Education backgrounds, Transaction history. These profiles were structured to match real-world segmentation logic used by retail banks.

2. **Organizational / Business Profiles**: Created for commercial banking use cases. Each profile includes: Company revenue (LLM-generated in realistic bands), Industry sector (Finance, Healthcare, Tech, etc.), Number of employees, Transaction history, Company growth. These were modelled to reflect B2B use cases for product recommendation.

3. **Transaction Histories**: Simulated based on inferred behaviour using GPT-generated rules: Categories: (Travel, Dining, Groceries, Utilities, etc.), Frequency & recency metrics, Max spend and monthly aggregates. These were key to identifying interests and deriving top 3 categories per user.

4. **Social Sentiment Data:** Customer comments were generated with varying tone, depth, and emotion using prompt-tuned LLM calls. We created both Retail and Commercial feedback threads covering: Complaints, Praise, Suggestions, Emotional intensity. This was used to train/test the sentiment-based offer engine.

5. **Product Eligibility & Rules:** Built a structured, tiered product catalogue (credit cards, debit cards, services) with Benefit mapping, Eligibility conditions (income, age, credit score).Backend Testing & Test Cases

We implemented a comprehensive set of unit and integration test cases for the backend FastAPI service:

- **Unit Tests** for individual ML components (churn model, recommender, sentiment integration)
- **Edge Case Tests** for missing data, corrupted inputs, and timeout simulations

LLM Prompt Engineering Strategy

Gemma 3 27B serves as the primary reasoning engine for HyperFin AI. To ensure consistency, safety, and relevance in financial contexts, we employed a structured approach to prompt engineering:
1) Prompts were framed using instruction-following templatessuch as: "Given the user's sentiment and transaction pattern, recommend a product with justification."
Financial terminology was embedded in the prompt context to guide Gemma's domain understanding.
2) Output formatting was constrained using few-shot examples to maintain response structure.
3) We implemented a (using logic probability) to detect low-certainty answers. In case of low confidence, or ambiguity, control was passed to fallback ML models (XGBoost or KNN).

Ethical Considerations

At the heart of HyperFin AI's design is a commitment to **ethical, responsible AI deployment**, especially critical in financial services. We proactively addressed fairness, transparency, privacy, and human oversight throughout our system.

1. Fairness & Bias Mitigation

- Sensitive attributes like **gender**, **religion**, and **ethnicity** were explicitly excluded from model inputs to avoid discriminatory outcomes.
- Recommendations were reviewed to prevent predatory targeting — e.g., avoiding loan suggestions to financially vulnerable users unless paired with financial literacy content or safeguards.
- Regular audits of model output distributions help identify and address potential indirect biases (e.g., age-based overfitting).

2. Explainability & Transparency

- Every model decision (especially for fraud alerts and churn risk) is accompanied by a **clear explanation** generated via SHAP values or natural language from the LLM.
- Fraud alerts include **human-readable justifications**, empowering users to understand why a flag was triggered.
- LLM responses are grounded in **verifiable reasoning**, with traceable prompt logic for accountability.

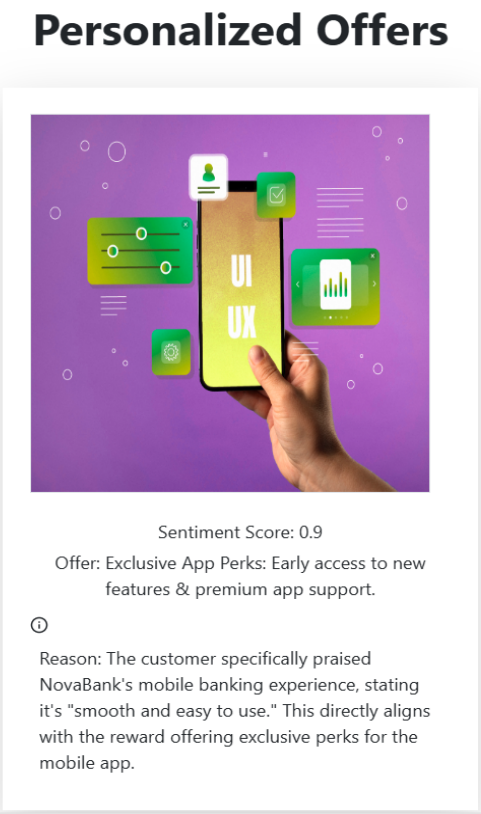
3. Privacy & Data Protection

- All datasets used are synthetic and privacy-safe. In real-world usage, HyperFin AI is designed to comply with data protection standards (e.g., GDPR, CCPA).
- Customer data is not stored or retained beyond the scope of real-time inference unless explicitly permitted.
- LLM prompts are anonymized and do not leak identifiable user data.

HyperFin AI is not just AI that performs — it is AI that performs ethically, transparently, and responsibly, in line with Nova Bank's customer-first values.

RESULTS

1. Personalized Offers



2. Recommended Products

Recommended Products

Nova Education Basic Credit Card

Why: The customer's PhD education and income meet the eligibility criteria. While their credit score is slightly below the Elite card's requirement, this card offers moderate benefits and is accessible with their score.

Nova Gourmet Dining Debit Card

Nova Everyday Shopping Credit Card

3. Banking Trivia

Banking Insights Just for You

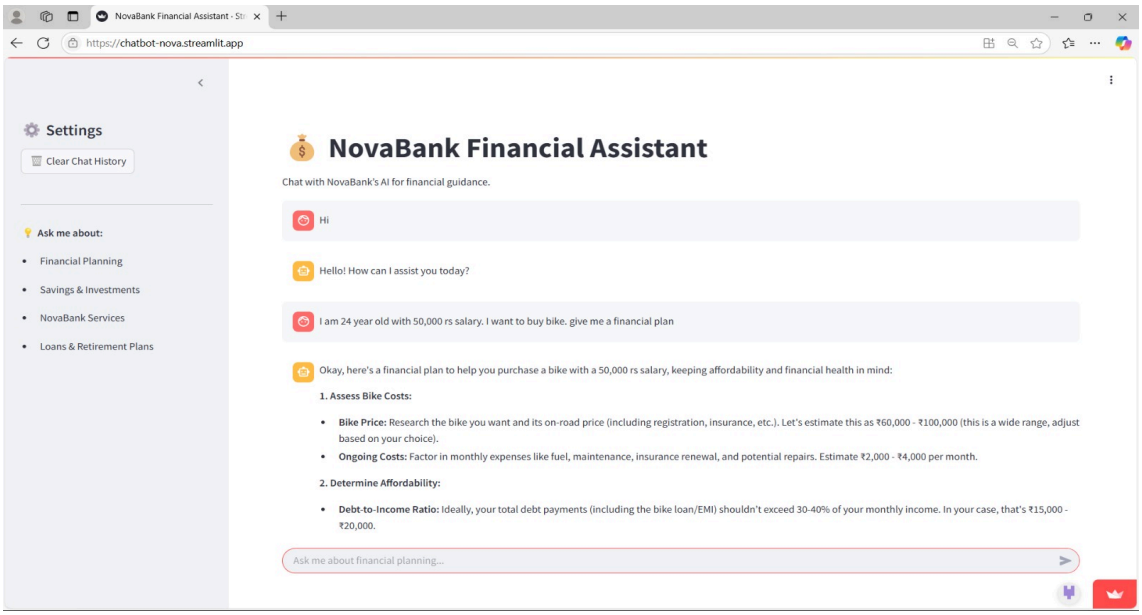
Comment: *Any advice on planning for retirement savings early?*

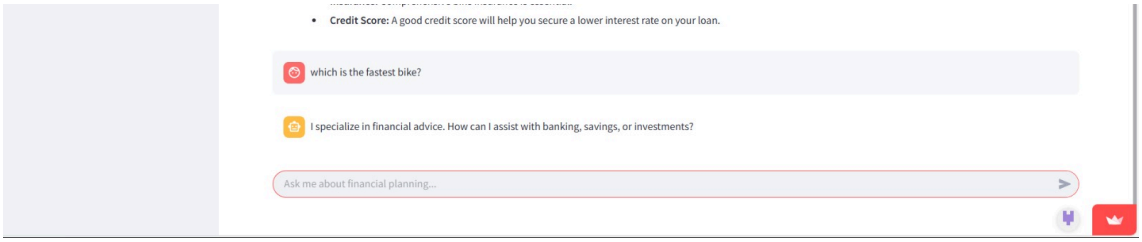
Platform: Twitter

Trivia: Starting retirement savings early leverages the power of compounding, allowing your investments to grow exponentially over time. Prioritize maximizing contributions to tax-advantaged accounts like 401(k)s or IRAs, even if starting with small amounts. Regularly review and adjust your investment strategy as you approach retirement to maintain a balanced and diversified portfolio.

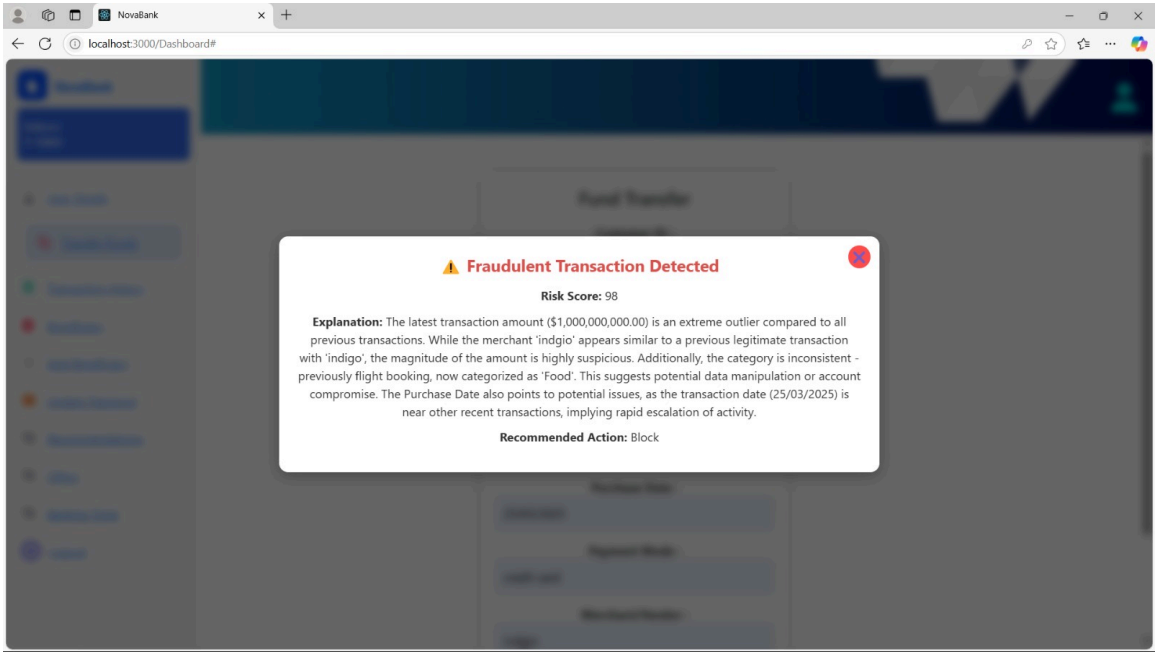
Recommended Product: NovaBank offers Individual Retirement Accounts (IRAs) with competitive interest rates and flexible contribution options. Our IRAs allow you to choose from a range of investment options to match your risk tolerance and time horizon. Opening an IRA with NovaBank provides a secure and straightforward way to begin building a comfortable retirement nest egg.

4. Banking Assistant





5. Fraud Detection



HyperFin AI demonstrates the powerful synergy between LLMs and traditional ML in solving real-world banking challenges. From enhancing product personalization to reducing churn and increasing customer trust, the system provides a scalable and explainable solution aligned with Nova Bank's customer-first philosophy. With modularity, real-time adaptability, and ethical safeguards built-in, HyperFin AI is both future-ready and practical for deployment. This project serves as a strong foundation for next-gen banking experiences powered by AI.