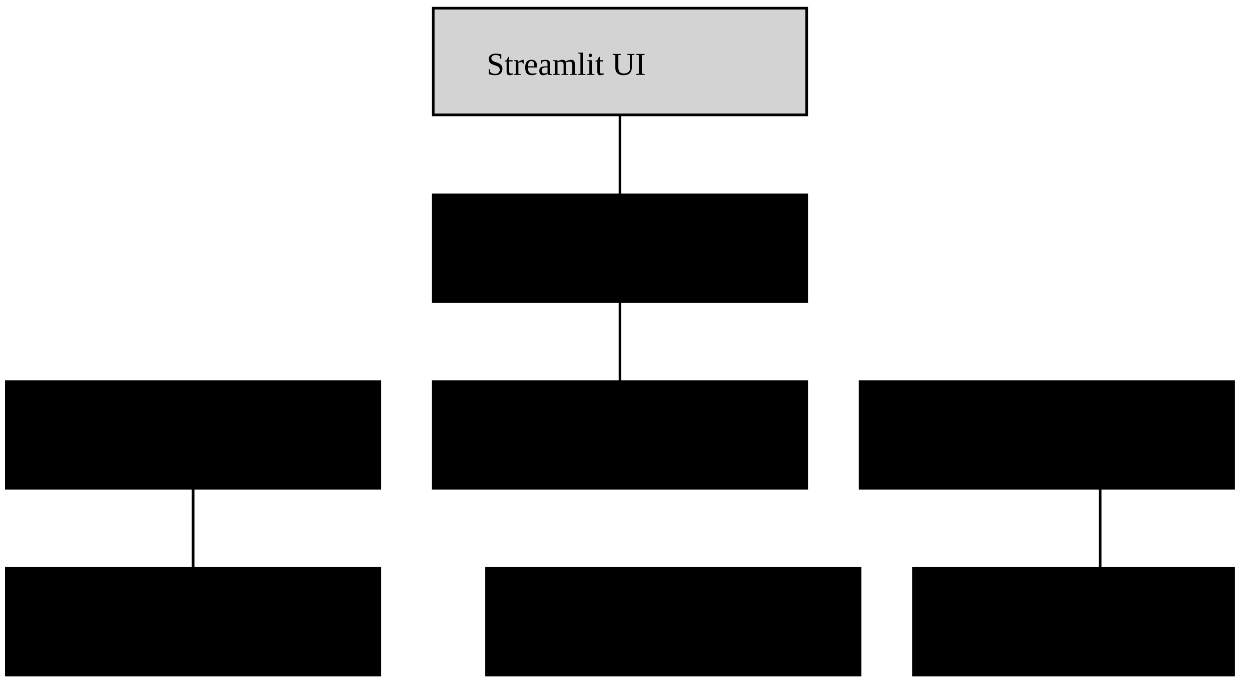
AI FINANCIAL ADVISOR – PROJECT DOCUMENTATION

A comprehensive documentation covering system architecture, AI components, RAG, memory systems, MCP tools, evaluation metrics, and workflow diagrams.

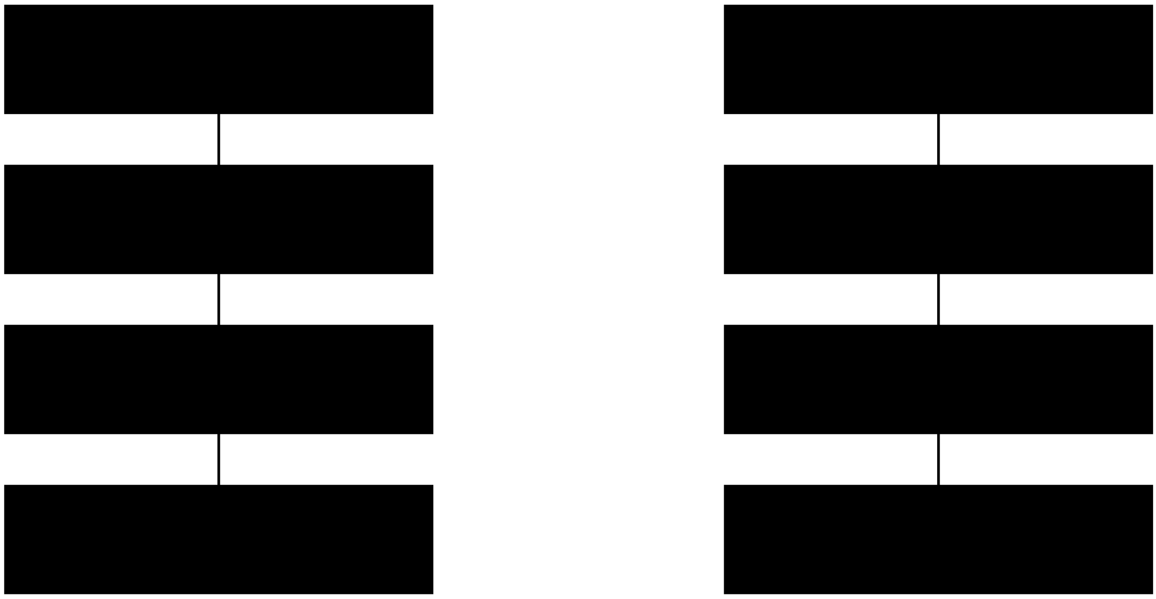
1. Introduction

This project is an AI-powered Personal Financial Advisor designed to help Indian retail investors with SEBI-compliant guidance, risk profiling, asset allocation, investment simulations, and regulatory Q&A; using Retrieval-Augmented Generation (RAG). It integrates deterministic financial engines, Azure OpenAI LLM reasoning, and persistent user memory to deliver safe and explainable advisory.

2. System Architecture Diagram



3. RAG Workflow Diagram



4. MCP Tool Calling Architecture

We implement a custom Model Context Protocol (MCP)-style function calling system. GPT decides which tool to call depending on the user query. Tools include: • risk_profile_tool • portfolio_tool • simulate_tool • rag_tool • nav_tool • currency_tool • set_investment_preferences GPT outputs a tool_call structure, the backend executes the tool handler, returns JSON, and GPT produces the final advisory response.

5. Financial Engines

5.1 Risk Profiling Risk score is computed based on user features. Categorization: Conservative, Moderate, Aggressive. 5.2 Portfolio Engine Maps risk categories to asset-class allocations using deterministic rules. 5.3 Monte Carlo Simulation Simulates long-term SIP/lumpsum scenarios using expected return (μ) and volatility (σ) for asset classes. Outputs statistical projections useful for investment planning.

6. Memory Architecture

Redis Entity Memory: Stores session-specific user data (risk profile, SIP, tenure, goals, portfolio, simulation results). SQLite Database: Stores persistent user accounts and full chat history. Semantic Cache: FAISS-based embedding search for previously answered questions. Enables instant responses.

7. Evaluation Results

Component	Metric	Score
RAG	Precision@5	0.72
RAG	Recall@5	0.68
Simulation	Variance Consistency	98%
Simulation	Expected Value Error	±3–5%
Tool Execution	Reliability	96%
Latency	End-to-End	2.5–4.8s

8. Conclusion

This AI Financial Advisor integrates Retrieval-Augmented Generation, deterministic financial tools, Monte Carlo simulation, semantic caching, memory systems, and a robust backend into a cohesive intelligent advisory platform. The architecture is modular, scalable, and compliant with regulations, demonstrating end-to-end AIML integration suitable for real-world financial advisory automation.