Agentic AI Solution Architecture

This section details the technical AI architecture powering AIRIS, including the LLM orchestration strategy, retrieval-augmented generation (RAG) pipeline, model selection rationale, and governance instrumentation.

Architecture Overview

AIRIS employs a **Multi-Agent Ensemble Architecture** combining specialized machine learning models for quantitative scoring with large language models (LLMs) for synthesis, explanation, and scenario narrative generation. The system integrates retrieval-augmented generation (RAG) to ground all outputs in authoritative International Bank (name changed), IMF, and ESG data sources.

Core Architectural Components:

1. **Scoring Engine Layer:** Ensemble of specialized ML models (XGBoost, Bayesian Networks) for ROI prediction, risk assessment, and ESG scoring
2. **RAG Pipeline:** Document retrieval and context injection from International Bank knowledge bases, IMF datasets, and ESG reporting frameworks
3. **LLM Synthesis Layer:** Foundation model for generating human-readable explanations, scenario narratives, and comparative analyses
4. **Governance Instrumentation:** Explainability tooling (SHAP, LIME), bias monitoring, and audit logging integrated at every inference point

LLM Selection & Orchestration

Foundation Model Selection

The LLM layer requires a model optimized for factual accuracy, long-context reasoning, and controlled output generation. Given International Bank security requirements and the need for on-premise or private cloud deployment, the recommended approach is:

Option	Model	Rationale
Primary (Recommended)	Claude 3.5 Sonnet (via AWS Bedrock) or GPT-4 Turbo (via Azure OpenAI)	Enterprise-grade security, 128K+ context window, strong reasoning capabilities, SOC 2 compliance
Alternative (On-Premise)	Llama 3.1 70B (fine-tuned) or Mixtral 8x22B	Full data sovereignty, no external API calls, customizable for domain-specific terminology

Orchestration Framework: LangChain or LlamaIndex serves as the orchestration layer, managing prompt routing, context injection from RAG, structured output parsing, and multi-step reasoning chains. The framework handles:

- Sequential reasoning chains for complex country assessments
- Parallel agent execution for ROI, Risk, and ESG evaluations
- Structured JSON output enforcement for dashboard integration
- Fallback and retry logic for API reliability

Retrieval-Augmented Generation (RAG) Pipeline

RAG ensures all LLM-generated explanations and recommendations are grounded in authoritative source documents, reducing hallucination risk and enabling citation-backed outputs.

Document Corpus

- **International Bank Open Data:** Country economic indicators, project evaluations, development reports
- **IMF Datasets:** Macroeconomic forecasts, fiscal sustainability assessments, Article IV reports
- **ESG Frameworks:** SASB standards, GRI indicators, International Bank Environmental and Social Framework documents
- **Internal Knowledge Base:** Historical investment decisions, post-implementation reviews, analyst notes

Embedding & Vector Storage

Component	Specification
Embedding Model	text-embedding-3-large (OpenAI) or BGE-large-en-v1.5 for on-premise deployment; 1536-3072 dimensions
Vector Database	Pinecone (managed) or pgvector on PostgreSQL (self-hosted) with HNSW indexing for sub-100ms retrieval
Chunking Strategy	Semantic chunking (500-1000 tokens) with 20% overlap; hierarchical chunking for structured reports preserving section context
Retrieval Method	Hybrid search (dense + sparse BM25) with cross-encoder reranking (ms-marco-MiniLM) for precision; top-k=10, reranked to top-5

Citation Enforcement: Every RAG-grounded output includes source attribution with document ID, section reference, and retrieval confidence score. The LLM prompt template enforces citation format: "[Source: {doc_title}, Section: {section}, Confidence: {score}]"

Prompt Engineering Strategy

Structured prompt templates ensure consistent, auditable outputs across all AIRIS functions. Each prompt includes role definition, context injection, output format specification, and ethical constraints.

Prompt Template: Investment Recommendation Explanation

```
[SYSTEM]
You are an expert financial analyst at the International Bank. Generate
clear, evidence-based explanations for investment recommendations. Always
cite specific data sources. Never speculate beyond provided data. Flag
uncertainty explicitly.
[CONTEXT]
Country: {country_name} | ROI Score: {roi_score} | Risk Score: {risk_score} |
ESG Score: {esg_score}
Retrieved Context: {rag_context}
[OUTPUT FORMAT]
Provide: 1) Executive Summary (2-3 sentences), 2) Key Factors (bullet list
with citations), 3) Risk Considerations, 4) Recommended Action, 5) Confidence
Level (High/Medium/Low with justification)
```

Guardrails & Hallucination Prevention

1. **Grounding Enforcement:** LLM outputs are validated against RAG-retrieved content; claims without source support are flagged for human review
2. **Structured Output Validation:** JSON schema enforcement ensures all outputs conform to expected structure; malformed outputs trigger regeneration
3. **Confidence Thresholds:** Recommendations below 70% model confidence require explicit "Low Confidence" labeling and mandatory human review
4. **Factual Consistency Checks:** Cross-reference numerical claims against source data; flag discrepancies exceeding 5% tolerance
5. **Human-in-the-Loop Gates:** All investment recommendations require analyst approval before presentation to governance boards

Model Governance & Explainability Integration

Governance tooling is integrated at the inference layer, not as a post-hoc audit:

- **SHAP Integration:** Feature importance computed for every ML model prediction; visualized in analyst dashboard
- **LIME Explanations:** Local interpretable explanations generated for individual country assessments
- **Model Cards:** Each model (ROI, Risk, ESG) maintains a versioned Model Card documenting training data, performance metrics, known limitations, and fairness evaluations
- **Audit Logging:** Every inference logged with input hash, model version, RAG context IDs, output, and timestamp; 7-year retention per International Bank policy
- **Bias Monitoring:** Quarterly fairness audits across geographic regions; alert thresholds for score distribution drift exceeding 2 standard deviations

Infrastructure & Performance Requirements

Requirement	Specification
End-to-End Latency	< 8 seconds for full country assessment (retrieval + scoring + explanation)
Vector Retrieval	< 100ms p99 for hybrid search + reranking
Concurrent Users	Support 50+ simultaneous analysts with < 10% latency degradation
Deployment	Azure Government Cloud (FedRAMP High) or International Bank private cloud; no data egress to public endpoints
Model Versioning	MLflow for experiment tracking, model registry, and deployment pipelines

[End of AI Solution Architecture Section]