NebulaByte_AI_Strategy_Report

Title: NebulaByte Multi-Agent Architecture and AI Strategy Report - 2025

Section 1: The Adaptive Multi-Agent Paradigm

NebulaByte's core AI division operates on a modular, multi-agent architecture (MAA) designed for highly autonomous decision-making and data processing. The MAA is governed by a central 'Controller Agent' responsible for task decomposition, tool/agent selection, and final synthesis. This agent utilizes a sophisticated LLM-based routing logic that dynamically selects one or more specialized agents (Web Search, ArXiv, PDF RAG) based on the user's query intent. This approach ensures maximal token efficiency and latency reduction. This system represents the future of enterprise knowledge management.

Section 2: Retrieval Augmentation Mechanics

The PDF RAG Agent is the cornerstone of our internal knowledge system. It employs a FAISS vector store deployed on a private cloud instance for fast, efficient nearest-neighbor search. Our ingestion pipeline uses the **sentence-transformers 'all-MiniLM-L6-v2' model** for generating embeddings, citing its superior performance-to-size ratio. The key retrieval setting is configured for **k=5** most relevant text chunks per query, ensuring a broad context is passed to the generation step, while also testing the LLM's ability to filter irrelevant noise. We've found that a chunk size of 500 characters works optimally for our specific data structure.

Section 3: Future Optimizations

The Q3 roadmap includes migrating the embedding model to a domain-specific fine-tuned variant to improve semantic search accuracy for industrial financial data. We are also researching techniques for hypothetical document embedding (HyDE) to further enhance retrieval for highly abstract or complex user queries. The success of the current FastAPI deployment framework has paved the way for a fully containerized, microservices-based agent deployment in Q4.