NebulaByte_Tech_Research_Summary

Title: Technical Deep Dive: Embedding Model Selection & Performance Benchmarks

Executive Summary: This report details the evaluation of various sentence-transformer models for our RAG system's embedding layer, focusing on two key metrics: **semantic accuracy (mAP@10)** and **latency (ms per 1000 tokens)**.

Model Benchmarks (Internal Data):

* Model Name: all-MiniLM-L6-v2, Dimensionality: 384, mAP@10 Score: 0.82, Latency (ms/1000): 120, Conclusion: **Optimal Balance**

* Model Name: all-MiniLM-L12-v2, Dimensionality: 384, mAP@10 Score: 0.84, Latency (ms/1000): 195, Conclusion: Higher latency, minimal gain

* Model Name: GTE-small, Dimensionality: 768, mAP@10 Score: 0.86, Latency (ms/1000): 310, Conclusion: Highest accuracy, too slow for production

Final Decision: We selected **'all-MiniLM-L6-v2'** as the default embedding model due to its optimal balance of retrieval quality (0.82) and speed. This choice directly impacts the system's Response Time metric, ensuring most queries are answered in under 3 seconds.

Future Research: We are investigating the deployment of a **smaller, custom-distilled model** trained specifically on legal and technical documentation. Initial results suggest this could increase semantic accuracy to **0.88** for domain-specific queries without significantly impacting latency. This upgrade is tied to the Q4 production cycle, starting November 1st.