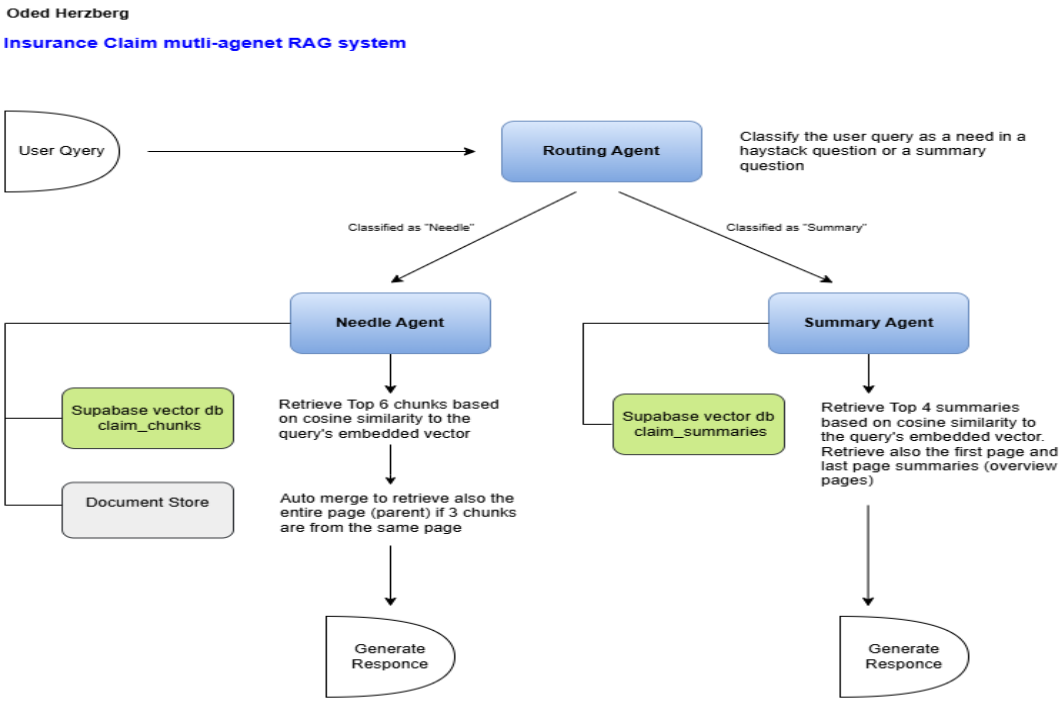
Insurance Claim Information Retrieval System

Multi-Agent RAG System with Specialized Retrieval Strategies

System Overview

This system implements a multi-agent RAG architecture for answering questions about insurance claims. A routing agent classifies incoming queries and dispatches them to specialized agents: a Needle Agent retrieves precise facts using small chunks (300 chars) with auto-merge for context expansion, while a Summary Agent handles high-level questions using page-level summaries (generated via MapReduce strategy) with always-included overview pages. Both agents leverage rich metadata (dates, involved parties, page types) to enhance retrieval accuracy. The system uses OpenAI GPT-4o-mini for generation, OpenAI embeddings for vector search, and Supabase (PostgreSQL + pgvector) for storage. Evaluation uses RAGAS with Google Gemini as an independent LLM judge.

System Architecture



Evaluation Results (RAGAS Framework)

Metric	Overall	Needle Agent	Summary Agent
Context Precision	0.534	0.607	0.461
Context Recall	0.605	0.800	0.410
Faithfulness	0.842	0.800	0.883
Answer Relevancy	0.830	0.782	0.879
Answer Similarity	0.940	0.949	0.932
Answer Correctness	0.687	0.837	0.538
Average Score	**0.740**	**0.796**	**0.684**

Key Findings: The Needle Agent achieves strong precision (0.949 similarity, 0.837 correctness) for specific factual queries. The Summary Agent excels at faithfulness (0.883) and relevancy (0.879), demonstrating effective synthesis of high-level information. Overall system performance of 0.740 indicates reliable retrieval and generation across diverse query types.