PROJECT CASE STUDY / OCTOBER 10, 2026

Multilingual RAG Chatbot | Retrieval, Streaming & Safety

Zahidul Islam · AI & Automation Engineer

Production assistant · human judgment required

Workflow and contribution

Paid client work on a multilingual health assistant: hybrid retrieval, reranking, emergency/crisis handling, structured grounding, SSE streaming and Redis caching. Web chat and controlled WhatsApp routing share the platform. Later work separates user-authored memory from generated replies and improves retrieval and image-only handling. August records include 28 passing backend test files. Full latency targets remain unmet; clinical content and final tone require human review.

Implementation and delivery boundaries

Paid client engineering for a multilingual health assistant: hybrid dense and keyword retrieval, cross-encoder reranking, emergency/crisis handling, structured grounding checks, SSE streaming and Redis caching. Web chat and controlled WhatsApp routing connect to the platform. Later work separates user-authored memory from generated replies and improves multilingual retrieval and image-only handling. The August release record includes 28 passing backend test files and a nine-check namespace migration. The recorded full latency target remains unmet; clinical content and final tone require human review. Historical deployment drift remains documented, so these records do not establish current source-to-live parity.

Tools

Node.js, Hybrid RAG, Pinecone

Evidence

August release records · October 10 public case study

This case study separates recorded verification from unfinished integrations and planned work. Historical checks describe their recorded release; they do not establish present service health, clinical acceptance, compliance certification or commercial capacity.