Component	Technology	Role
Live Inference	Groq (Llama 3.3 70B)	The user-facing brain. Selected for 0.2s latency to impress clients.
Embeddings	Together AI	Model: m2-bert-80M-8k-retrieval. CRITICAL: Used in both Backend and Frontend to ensure vectors match.
Live Knowledge	Hugging Face API (Tool)	Gives the agent real-time access to the latest models (e.g., "What is the best model for medical coding today?").
Deep Knowledge	Redis Stack (Cloud)	Stores your RAG vectors + Semantic Cache.
Orchestration	Vercel AI SDK (Core)	Manages the chat stream, tool calling (Hugging Face), and Zod validation (Guardrails).
Automation	LangChain (Python)	The heavy lifter. Runs on GitHub Actions to scrape and process data.
Summarization	Google Vertex AI	Used ONLY in the backend script to summarize long articles cheaply (using your credits).