



RAG-Based Agent AI Chatbot — Cloudless Knowledge Assistant

This project is a fully self-contained Retrieval-Augmented Generation (RAG) based agent chatbot, designed to function entirely without cloud dependencies. It only responds with data it has in its internal knowledge base, enabling use cases like:

 Personal second-brain / note-based assistant

 Contextualized customer support

 Personalized learning or career FAQ bots

Key Features

Fully Local or Cloudless Operation: All data is stored and queried locally using pgvector with NeonDB — no OpenAI-hosted vector store.

Next.js 14 Frontend using App Router and Vercel AI SDK

Drizzle ORM for type-safe SQL generation and migrations

OpenAI Embeddings (or local fallback embeddings with Nomic/Ollama)


Shadcn UI + TailwindCSS v4 for the interface and theming

Agentic Behavior: The chatbot can store facts during a conversation and recall them in future exchanges, mimicking memory.

Architecture Summary

 Embedding Engine: OpenAI's text-embedding-3-small for fast semantic representation

 Vector Store: pgvector extension inside NeonDB (Postgres)

 ORM: Drizzle used to interface with vectorized DB entries via SQL

 UI: Vercel AI SDK handles input/output streaming in real-time