THIA: NORC's Trusted Health Information Assistant

Technical Documentation

THIA is an agentic RAG platform that delivers trusted, evidence-based health information through an interactive AI interface. It uses a React frontend, FastAPI backend, PostgreSQL with pgvector, and AWS Bedrock to ingest, search, and cite curated sources, ensuring responses are grounded in verifiable evidence rather than generic chatbot output. The system is architected on AWS with a layered design that keeps infrastructure concerns cleanly separated — built for security, scalability, and deployment flexibility.

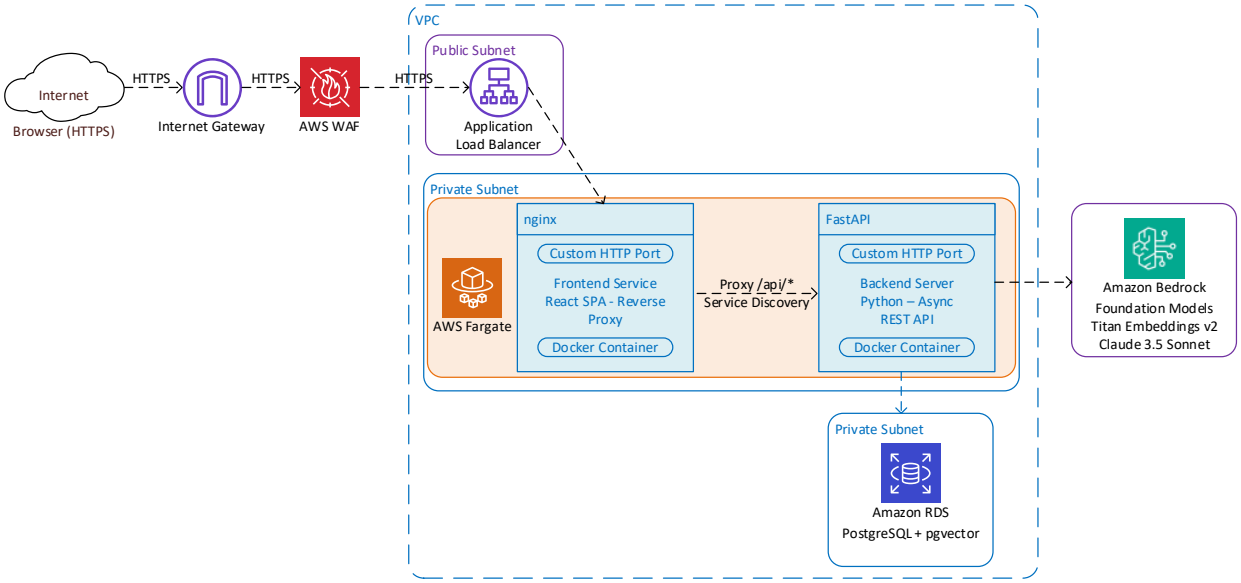
AWS-Powered Architecture

THIA runs on a purpose-built AWS cloud architecture designed to keep public traffic, application services, database storage, and managed AI services cleanly separated. This layered design enables reliable performance, secure data handling, and the modular flexibility needed to support diverse health communication use cases. The components below work together to power THIA's full pipeline — from content ingestion to grounded, citation-backed response generation.

Infrastructure Components	Agentic RAG Pipeline
• **AWS WAF + ALB** handle HTTPS entry, filtering, SSL termination, and routing. • **ECS/Fargate** runs containerized frontend and backend services serverlessly. • **nginx** serves the React SPA and proxies API requests. • **FastAPI** exposes authenticated APIs for projects, sources, ingestion, RAG indexing, search, and chat. • **Amazon RDS PostgreSQL** (pgvector) stores data, chunks, metadata, and embeddings. • **Amazon Bedrock** (Titan Embeddings v2, Claude Sonnet) powers vectorization, enrichment, and answer generation.	1. **Source ingestion:** Crawl4AI fetches web content as RAG-ready markdown; PDF text extraction is also supported. 2. **Content cleanup:** Markdown undergoes rule-based processing, structure normalization, and optional Claude Sonnet cleanup via AWS Bedrock. 3. **Semantic chunking:** Content is split into token-aware chunks using a sliding-window strategy with overlap, preserving document structure. 4. **Metadata enrichment:** Titles, authors, keywords, topics, page types, word counts, and relevance scores are extracted to support filtering and quality review. 5. **Vector indexing:** Amazon Titan Embeddings v2 generates 1024-dimensional embeddings stored in PostgreSQL via pgvector. 6. **Grounded retrieval and response:** User questions are embedded, matched to relevant chunks, and passed to Claude Sonnet with source context; responses include citation markers tied to retrieved sources.

The infrastructure components power the six-stage RAG pipeline that transforms curated web and document sources into searchable, citable knowledge. This approach helps organizations move from static health content to conversational access while keeping answers grounded in a managed knowledge base. See **Figure 1**.

Figure 1. THIA Architecture



Modern Cloud-Native Application Stack

THIA is built on a modern, cloud-native stack in which each layer has a defined role—from the browser-facing interface to the AI services that power retrieval and response. The table below maps each component to its function within the system.

Layer	Technologies	Role in THIA
Frontend	React, Vite, Tailwind CSS, Axios, Framer Motion, nginx	Browser application, dashboard, source management, search, chat experience
Backend	FastAPI, Pydantic, SQLAlchemy, asyncpg, Uvicorn/Gunicorn	API services for authentication, projects, scraping, processing, RAG, chat
Data	Amazon RDS PostgreSQL, pgvector	Application database, source/chunk storage, vector search, retrieval metadata
AI services	AWS Bedrock, Amazon Titan Embeddings v2, Claude Sonnet LLMs	Embedding generation, markdown cleanup, metadata extraction, relevance scoring, response generation
Ingestion	Crawl4AI, PyPDF2, tiktoken	Web crawling, PDF extraction, markdown conversion, token-aware processing
Deployment	Docker, nginx, ECS/Fargate, Application Load Balancer, AWS WAF	Containerized deployment path with AWS-managed routing/protection
Evaluation	Promptfoo	RAG quality checks, citation checks, red-team tests, adversarial validation

Partner with NORC

NORC brings deep expertise in public health, data science, and health communication, with THIA as a core capability for delivering trusted, AI-enabled health information to patients and consumers. Built by Mehmet Celepkolu, PhD, a nationally recognized expert in NLP, machine learning, and advanced data analytics, every deployment is scientifically grounded, ethically responsible, and aligned with best practices for AI in health communication. Our interdisciplinary team supports organizations across the full lifecycle—from knowledge base development and AI customization to deployment, UX design, and ongoing optimization—producing a scalable, interactive platform that extends reach while giving people access to information they can trust. To discuss deployments, demonstrations, or partnerships, contact Dr. Celepkolu at **celepkolu-mehmet@norc.org**.

> *"THIA was built to do something deceptively hard: deliver health information that is both genuinely trustworthy and accessible. The architecture reflects that goal at every layer, from knowledge ingestion and structuring to response generation and citation." — **Mehmet Celepkolu**, Principal Data Scientist & Lead Developer, THIA*