Chat-bot medical data backend architecture documentation

**Overview:**

This backend system is designed to process medical book data, create semantic embeddings for efficient knowledge retrieval, and serve query results through a large language model (LLM). The system leverages LangChain for orchestration, Pinecone as a vector database (VDB), and supports multiple LLM options, including OpenAI and locally hosted models such as Ollama. Additionally, it integrates a custom DeepSeek configuration for enhanced document extraction and chaining.

**System Workflow:**

1. Data Ingestion and Processing

Source: Medical Book (raw data)

Step 1: Extract all documents from the medical book.

Step 2: Split the extracted content into three separate chunks for manageable processing:

* + - * Chunk 1
      * Chunk 2
      * Chunk 3

1. Embedding Generation

Each chunk is individually converted into a vector embedding:

* + - Vector 1 from Chunk 1
    - Vector 2 from Chunk 2
    - Vector 3 from Chunk 3

1. Semantic Index Construction
   * + The three vector embeddings are combined to build a semantic index.
     + This semantic index is stored and managed in the Pinecone Vector Database (VDB), serving as the knowledge base for efficient similarity search.
2. Query Handling
   * + When a user submits a query:
       1. The query text is converted into a vector embedding.
       2. This vector is used to query the Pinecone knowledge base.
       3. The knowledge base returns ranked results based on similarity scores.
       4. The results are passed to a Large Language Model (LLM) for generating the final response.

**# Backend Architecture**

Medical Book

Chunk 1

Chunk 2

Extract All Docs

Chunk 3

Embedding models

Vector 3

Vector 2

Vector 1

Semantic index

Knowledge Base using Pinecone VDB

**# Frontend Architecture**

User Input Query

Query embedding text converted into

Vector embedding

It goes into

Knowledge Base

Rank Results

I’ll be using one large language model to show

**Technology stack:**

|  |  |
| --- | --- |
| **Component** | **Description** |
| OpenAI LLM | |  | | --- | | Cloud-based LLM for fast, reliable inference and response generation. |  |  | | --- | |  | |
| LangChain | Framework to orchestrate LLMs and vector search components. |
| Pinecone | Vector database for managing semantic embeddings and fast similarity search. |
| Flask | Backend web framework serving API endpoints and integrating components. |
| Open source LLM (optional) | Locally hosted LLM alternative, but performance depends heavily on machine configuration. |
| Ollama Local LLM | Locally hosted language model managed via Ollama, suitable for systems with sufficient resources. |
| Custom DeepSeek Configuration | Cloud-based LLM for fast, reliable inference and response generation. Specialized document extraction and chunking setup to optimize data ingestion from medical texts. |

**Local LLM Configuration: Ollama**

* Ollama allows you to run open-source language models locally on your machine.
* System requirements:
  + High-performance CPU and GPU recommended.
  + Sufficient RAM and storage for model hosting.
* Performance Notes:
  + Local inference time can vary significantly depending on hardware.
  + Ollama enables offline usage and data privacy but may have slower responses compared to cloud based LLMs.

**OpenAI API Pricing**

OpenAI offers a range of models with varying capabilities and pricing. Below are some of the prominent models.

GPT-4.5

* Context length – 128k tokens
* Input - $75.00 per 1M tokens
* Cached Input - $37.50 per 1M tokens
* Output - $150.00 per 1M tokens

GPT-4.1

* Input: $2.00 per 1M tokens
* Cached Input: $0.50 per 1M tokens
* Output: $8.00 per 1M tokens

GPT-4.1 Mini

* Input: $0.40 per 1M tokens
* Cached Input: $0.10 per 1M tokens
* Output: $1.60 per 1M tokens

GPT-4.1 Nano

* Input: $0.10 per 1M tokens
* Cached Input: $0.025 per 1M tokens
* Output: $0.40 per 1M tokens

**DeepSeek API Pricing**

DeepSeek provides cost-effective models suitable for various applications. Pricing is based on token usage and varies depending on the time of day due to off-peak discounts.

Standard Pricing (UTC 00:30–16:30)

* **deepseek-chat**:
  + Input (Cache Hit): $0.07 per 1M tokens
  + Input (Cache Miss): $0.27 per 1M tokens
  + Output: $1.10 per 1M tokens
* **deepseek-reasoner**:
  + Input (Cache Hit): $0.14 per 1M tokens
  + Input (Cache Miss): $0.55 per 1M tokens
  + Output: $2.19 per 1M tokens

**Pinecone Vector Database Pricing**

Pinecone offers scalable vector database solutions with different pricing tiers:

Starter Plan

* **Cost**: Free
* **Features**:
  + Up to 2 GB storage
  + Community support
  + Access to serverless indexes

Standard Plan

* **Cost**: Starting at $25/month
* **Includes**:
  + $15/month usage credits
  + Pay-as-you-go for serverless, inference, and assistant usage
  + Multiple projects and users
  + Backup and restore
  + Prometheus metrics
  + Free support

Enterprise Plan

* **Cost**: Starting at $500/month
* **Includes**:
  + $150/month usage credits
  + Everything in Standard Plan
  + 99.95% uptime SLA
  + SAML SSO
  + Private networking
  + Customer-managed encryption keys
  + Audit logs
  + HIPAA compliance
  + Pro support

For more detailed and up-to-date information, please refer to the official pricing pages:

* OpenAI: <https://openai.com/api/pricing/>
* DeepSeek: <https://api-docs.deepseek.com/quick_start/pricing>
* Pinecone: <https://www.pinecone.io/pricing/>