

# Gemini Chat

I design a general three-layer architecture:

1. LLM Provider Layer:

- Including Ollama, vLLM, llama.cpp
- Unified OpenAI Compatible APIs

2. Services Layer

- CoreLogicService
- InputDataHandleService
- PromptService
- RetrievalService
- StateTransitionService

3. Presentation Layer

- Optimistic Progressive Markdown Parsing
- Data Export

now I need you to integrate the three-layer architecture to a python+FastAPI web application, which can let users to upload files and chat with the data in these files. It means the system need to extract, analyze and chunk data from these files. The most important is system's answers only and always from there uploaded files. Architecting my three-layer architectures into the general python+FastAPI web application architecture as you are a talent and proficient architect

- ◆ 這是一個非常好的三層架構設計，完全符合現代 RAG (Retrieval-Augmented Generation) 系統的設計理念。將您的架構整合到一個健壯的 Python + FastAPI 應用程式中，關鍵在於實現「關注點分離」(Separation of Concerns) 和「依賴注入」(Dependency Injection)。

作為您的架構師，我將把您的三層模型映射到一個可維護、可擴展的 FastAPI 專案結構中。

## 總體架構藍圖：RAG 流程整合

您的三層架構將被整合成兩個核心工作流程：**1. 數據提取 (Ingestion)** 和 **2. 聊天檢索 (Retrieval & Chat)**。

### 1. 數據提取流程 (File Upload)

[使用者] -> [Presentation Layer (Client)] -> FastAPI Endpoint (/upload) -> [Services Layer (InputDataHandleService)] -> [Services Layer (RetrievalService)] -> [Vector DB]

1. **使用者** 上傳一個或多個文件。
2. **FastAPI Endpoint** 接收 `UploadFile` 對象。
3. **InputDataHandleService** 負責：
  - 從文件中提取原始文本（例如使用 `pypdf` 處理 PDF）。
  - 將文本進行「分塊」(Chunking)。
  - 調用 `RetrievalService`。
4. **RetrievalService** 負責：
  - 為每個文本塊 (chunk) 產生嵌入 (Embeddings)。(這通常是透過一個嵌入模型完成，它可以是您 LLM Provider 的一部分，或是一個獨立的模型如 `sentence-transformers`)。
  - 將文本塊及其嵌入向量存儲到 **Vector Database**（例如 ChromaDB, FAISS, Milvus）。

### 2. 聊天檢索流程 (Chat)



```
├─ requirements.txt
├─ uvicorn_runner.py                # 用於啟動 uvicorn
```

## 整合您的三層架構

以下是將您的每一層具體實現到上述結構中的方法。

### 1. LLM Provider Layer (實現)

這一層在 FastAPI 專案中表現為一個**客戶端 (Client)**。我們在 `app/llm_provider/client.py` 中定義它。

- **統一 OpenAI 兼容 API**：這是關鍵。無論您後端是 Ollama、vLLM 還是 llama.cpp，它們都提供了 OpenAI 兼容的 `/v1/chat/completions` 端點。
- **實現**：我們將創建一個可重用的服務，它從 `app/core/config.py` 讀取 `LLM_PROVIDER_BASE_URL`（例如 `http://localhost:11434/v1` 或 `http://localhost:8000/v1`）。

`app/llm_provider/client.py` (範例):

Python



```
import httpx
from app.core.config import settings # 從 Pydantic SettingsManagement 讀取

class LLMPProviderClient:
    def __init__(self):
        # 從 .env 讀取 LLM 服務的基礎 URL
        self.base_url = settings.LLM_PROVIDER_BASE_URL
        self.api_key = settings.LLM_PROVIDER_API_KEY or "ollama" # Ollama/vLLM 通常不需要 key

    async def get_chat_completion_stream(self, messages: list[dict], model: str = "default-model"):
        """
        請求 LLM 進行流式聊天補全
        """
        async with httpx.AsyncClient(timeout=60.0) as client:
            request_body = {
                "model": model,
                "messages": messages,
                "stream": True,
                # 這裡可以添加其他參數, 如 temperature=0.0
            }

            headers = {"Authorization": f"Bearer {self.api_key}"}

            async with client.stream(
                "POST",
                f"{self.base_url}/chat/completions",
                json=request_body,
                headers=headers
            ) as response:
                response.raise_for_status() # 確保請求成功
                async for chunk in response.aiter_bytes():
                    yield chunk # 直接將原始 SSE 數據塊傳遞出去
```

### 2. Services Layer (實現)

這是您應用的**大腦**。我們使用 FastAPI 的**依賴注入 (Dependency Injection)** 系統來管理這些服務。

`app/services/retrieval_service.py` (範例):

Python



```
from some_vector_db import VectorStore # 假設使用 ChromaDB 或 FAISS
from some_embedding_model import EmbeddingModel # 假設使用 sentence-transformers
```

```

class RetrievalService:
    def __init__(self):
        # 理想情況下，這些應該是單例（Singleton）模式
        self.vector_store = VectorStore(path="./db/chroma_db")
        self.embedding_model = EmbeddingModel('all-MiniLM-L6-v2')

    async def add_chunks(self, file_id: str, chunks: list[str]):
        embeddings = self.embedding_model.embed(chunks)
        metadatas = [{"file_id": file_id, "chunk_index": i} for i, _ in enumerate(chunks)]
        self.vector_store.add(ids=[f"{file_id}_{i}" for i in range(len(chunks))],
                               documents=chunks,
                               embeddings=embeddings,
                               metadatas=metadatas)

    async def query(self, query_text: str, file_ids: list[str], top_k: int = 5) -> list[str]:
        query_embedding = self.embedding_model.embed([query_text])[0]

        # 關鍵：只在使用者上傳的文件中搜索
        results = self.vector_store.query(
            query_embeddings=[query_embedding],
            n_results=top_k,
            where={"file_id": {"$in": file_ids}} # 假設的過濾語法
        )
        return results['documents'][0] # 返回相關的文本塊

```

app/services/core\_logic\_service.py (範例):

Python



```

from fastapi import Depends
from app.llm_provider.client import LLMPProviderClient
from app.services.retrieval_service import RetrievalService
from app.services.prompt_service import PromptService
from app.services.state_transition_service import StateTransitionService

class CoreLogicService:
    def __init__(
        self,
        retrieval_service: RetrievalService = Depends(),
        prompt_service: PromptService = Depends(),
        llm_client: LLMPProviderClient = Depends(),
        state_service: StateTransitionService = Depends()
    ):
        # 透過 Depends(), FastAPI 會自動注入這些服務的實例
        self.retrieval_service = retrieval_service
        self.prompt_service = prompt_service
        self.llm_client = llm_client
        self.state_service = state_service

    async def generate_response_stream(self, user_id: str, chat_id: str, query: str):
        # 1. 獲取此聊天的文件範圍（假設已儲存）
        allowed_file_ids = await self.state_service.get_files_for_chat(chat_id)

        # 2. Retrieval: 檢索上下文
        context_chunks = await self.retrieval_service.query(query, allowed_file_ids)

        # 3. Prompt: 構建提示詞（這是實現「僅從文件回答」的關鍵）
        messages = self.prompt_service.build_rag_prompt(query, context_chunks)

        # 4. State Transition: (非同步) 保存用戶訊息
        await self.state_service.save_message(chat_id, "user", query)

        # 5. LLM Provider: 獲取流式回應
        # 我們將 LLM 的流直接傳遞回客戶端
        response_stream = self.llm_client.get_chat_completion_stream(messages)

        # 6. (可選) 在流結束後保存 AI 回應
        # full_response = await self.state_service.process_and_save_stream(chat_id, response_stream)

```

```
# return full_response

return response_stream
```

### 3. Presentation Layer (實現支持)

您的後端 (FastAPI) **不包含** Presentation Layer 本身（它是一個獨立的 Web 前端，例如 React/Vue）。您的後端職責是**提供 API 來支持這一層**。

- **Optimistic Progressive Markdown Parsing**：為了支持這一點，您的 `/chat` 端點**必須**使用 `StreamingResponse`。

`app/api/v1/endpoints/chat.py` (範例):

Python



```
from fastapi import APIRouter, Depends, UploadFile, File
from fastapi.responses import StreamingResponse
from app.services.core_logic_service import CoreLogicService
from app.services.input_data_handle_service import InputDataHandleService
from app.models.schemas import ChatRequest

router = APIRouter()

@router.post("/chat")
async def chat_endpoint(
    request: ChatRequest,
    core_service: CoreLogicService = Depends()
):
    """
    接收用戶查詢並以流式回應
    """
    stream = await core_service.generate_response_stream(
        user_id=request.user_id,
        chat_id=request.chat_id,
        query=request.query
    )

    # 關鍵：將 LLM 服務的流直接轉發給客戶端
    # 這使得 Markdown 能夠漸進式解析
    return StreamingResponse(stream, media_type="text/event-stream")

@router.post("/upload")
async def upload_files(
    files: list[UploadFile] = File(...),
    input_service: InputDataHandleService = Depends()
):
    """
    處理文件上傳、提取和分塊
    """
    file_ids = []
    for file in files:
        file_id = await input_service.process_and_store_file(file)
        file_ids.append(file_id)

    return {"message": "Files processed successfully", "file_ids": file_ids}
```

- **Data Export**：這只是 `StateTransitionService` 的一個簡單 API 端點，用於查詢數據庫並將聊天記錄格式化為 JSON 或 CSV。

---

**關鍵約束：「僅從上傳的文件中回答」**

這是整個架構中**最重要**的業務規則。它**不是**透過 LLM Provider 層（例如 vLLM）的配置來實現的，而是透過**服務層 (Services Layer)** 中的 `PromptService` 來**嚴格執行**的。

您的 `PromptService` 必須產生一個類似這樣的提示詞：

`app/services/prompt_service.py` (範例模板):

Python

```
class PromptService:
    def __init__(self):
        self.SYSTEM_PROMPT_TEMPLATE = """
        您是一個嚴謹的問答機器人。
        您必須「僅」使用下面提供的「上下文」來回答使用者的問題。
        絕對禁止使用您的任何內部知識或來自上下文之外的資訊。

        如果「上下文」中沒有足夠的資訊來回答問題，您必須明確回答：
        「根據您提供的文件，我找不到相關資訊。」

        請勿杜撰、猜測或推斷「上下文」中未明確說明的內容。

        ---
        [上下文]
        {context}
        ---
        """

    def build_rag_prompt(self, query: str, context_chunks: list[str]) -> list[dict]:
        # 將所有上下文塊合併為一個字符串
        context_str = "\n\n".join(context_chunks)

        system_prompt = self.SYSTEM_PROMPT_TEMPLATE.format(context=context_str)

        # 遵循 OpenAI 的消息格式
        return [
            {"role": "system", "content": system_prompt},
            {"role": "user", "content": query}
        ]
```

透過這種方式，您不是在**請求** LLM 合作，而是在**指示**它在一個嚴格受限的沙盒（即您提供的上下文）內運作。這就是您實現「答案**僅**來自上傳文件」的核心機制。

Presentation Layer must implement using optimistic progress markdown parsing(OPMP) design

◆ 您說的沒錯。**Optimistic Progressive Markdown Parsing (OPMP)** 是現代聊天機器人 UI 中最關鍵的使用者體驗 (UX) 設計之一，它提供了「打字機」般的即時反饋感。

在我們的三層架構中，這個需求**完全**由 **Presentation Layer (前端)** 實現，而您的 **FastAPI 後端 (Services Layer + LLM Layer 接口)** 則扮演**啟用者 (Enabler)** 的角色。

這兩層的職責劃分非常明確：

## 1. 後端 (FastAPI) 的架構支持：啟用 OPMP

您的後端架構**不需要**解析 Markdown。它的**唯一且關鍵**的職責是：以**最低的延遲**，將 LLM Provider 層的令牌流 (token stream) **原封不動地、即時地傳遞給前端**。

您在 `chat.py` 端點中使用的 `StreamingResponse` 正是實現此目的的完美工具。

Python

```
# app/api/v1/endpoints/chat.py (回顧)

@router.post("/chat")
async def chat_endpoint(
    request: ChatRequest,
    core_service: CoreLogicService = Depends()
):
    # 1. core_service 調用 LLM Provider (Ollama/vLLM)
    # 並獲取一個異步生成器 (async generator)
    stream = await core_service.generate_response_stream(...)

    # 2. 關鍵：
    # FastAPI 不緩衝 (buffer) 回應。
    # LLM 產生一個 token，FastAPI 就將該 token
    # 封裝成一個 SSE (Server-Sent Events) 封包並立即發送。
    return StreamingResponse(stream, media_type="text/event-stream")
```

架構上的含義：

- **低延遲**：這是 OPMP 的前提。您的後端是一個「透明代理」，幾乎不增加流式傳輸的開銷。
- **協議**：`text/event-stream` (Server-Sent Events) 是前端實現 OPMP 的標準協議。前端將使用 `EventSource` API 或 `fetch` API 的 `ReadableStream` 來消費 (consume) 這個流。

## 2. 前端 (Presentation Layer) 的實現關鍵：實現 OPMP

您的前端 (例如 React, Vue, Svelte) 將接收這個 SSE 流。OPMP 的所有複雜邏輯都在這裡發生。

「Optimistic Progressive Markdown Parsing」可以分解為三個動作：

### A. Progressive (漸進式)

前端**不會**等待流結束。它會訂閱 SSE 流，每當一個新的數據塊 (chunk) 到達時 (例如 `data: {"content": "..."}` )，它就會立即處理。

JavaScript



```
// 範例：前端 JavaScript (使用 fetch)
const response = await fetch('/api/v1/chat', { ... });
const reader = response.body.getReader();
const decoder = new TextDecoder();
let fullResponse = "";

while (true) {
    const { done, value } = await reader.read();
    if (done) break;

    const chunk = decoder.decode(value);
    // (這裡還需要解析 SSE 格式，簡單起見先假設 chunk 是純文本)

    // 1. 漸進式地累積文本
    fullResponse += chunk;

    // 2. 樂觀地 (Optimistic) 渲染
    renderMarkdown(fullResponse); // <-- 關鍵步驟
}
```

### B. Optimistic (樂觀)

`renderMarkdown(fullResponse)` 這一步就是「樂觀」的體現。我們**假設** `fullResponse` 變數中不完整的 Markdown 字符串是可以被渲染的。

- **樂觀更新**：我們不等待一個完整的 Markdown 區塊 (例如整個代碼塊或列表) 完成，而是**立即**將當前的、不完整的 `fullResponse` 扔給 Markdown 解析器。

- **UI 狀態更新**：在 React 中，這意味著 `setState(prev => prev + newChunk)`。

### C. Markdown Parsing (Markdown 解析)

這是最棘手的部分。一個常規的 Markdown 解析器 (Parser) 在遇到不完整的語法時（例如 ````python` 但沒有結尾的 `````）可能會崩潰或渲染錯誤。

您的 OPMP 解析器**必須具備容錯性 (Fault Tolerance)**：

1. **渲染已知內容**：它會渲染所有已完成的 Markdown 語法（例如一個完整的段落）。
2. **處理不完整語法**：
  - **範例 1 (代碼塊)**：當解析器遇到 ````python\nprint("hi` 時，它應該**樂觀地**開始渲染一個代碼塊。它不會因為沒有結尾的 ````` 而拒絕渲染。
  - **範例 2 (列表)**：當解析器遇到 `* 項目一\n* 項` 時，它應該渲染「項目一」列表，並將「\* 項」暫時視為純文本，直到 `目二` 到達，然後**重新解析**並將其更正為第二個列表項。

**總結**：

您的三層架構設計是**完美契合** OPMP 需求的。

- **後端 (FastAPI)**：作為一個高效的、非緩衝的**流式數據提供者 (Stream Provider)**。
- **前端 (Presentation Layer)**：作為一個複雜的**流式數據消費者 (Stream Consumer)**，它內部包含一個能處理不完整輸入的、容錯的 Markdown 渲染引擎。