# 📦 RAG Pipeline - End-to-End Production Flow  
  
本專案實現了從 \*\*PDF 資料 → Chunk 切分 → 問題生成 → 作答 → 審核 → RAG / 微調資料輸出\*\* 的全流程。   
支援本地 LLM（如 Ollama / llama.cpp / vLLM）與雲端 OpenAI API。  
  
---  
  
## 🚀 環境準備  
  
### 1. Python 環境  
```bash  
python -m venv .venv  
source .venv/bin/activate # macOS/Linux  
.venv\Scripts\activate # Windows PowerShell  
pip install -r requirements.txt  
```  
  
### 2. Docker（可選）  
如需封裝整個流程，請安裝：  
- Docker 20+  
- Docker Compose 2+  
  
---  
  
## 📂 資料準備  
- 將 PDF 放入 `data/` 資料夾  
- 確認檔名可辨識（避免中文亂碼）  
  
---  
  
## 🧩 流程步驟  
  
### 1️⃣ 抽取與索引  
```bash  
python scripts/01\_ingest.py --pdf-dir data --out-dir indices  
```  
輸出：  
- `indices/chunks.txt`（BLOCK / JSONL 格式）  
- `indices/metadata.json`  
  
---  
  
### 2️⃣ 問題與關鍵字生成  
設定模型：  
```powershell  
$env:LLM\_MODE="OPENAI\_COMPAT"  
$env:OPENAI\_BASE\_URL="http://localhost:11434/v1"  
$env:OPENAI\_API\_KEY="sk-local"  
$env:MODEL\_QUESTION="llama3.2:latest"  
$env:MODEL\_KEYWORDS="llama3.2:latest"  
```  
執行：  
```bash  
python -m packages.rag.generation.Question\_llm.make\_questions --index indices --out outputs/questions.jsonl --langs zh,en  
```  
檢查：  
```bash  
python scripts/check\_questions.py outputs/questions.jsonl  
```  
  
---  
  
### 3️⃣ 作答  
```bash  
python -m packages.rag.generation.Answer\_llm.make\_answers --index indices --questions outputs/questions.jsonl --out outputs/answers.jsonl --max 100  
```  
  
---  
  
### 4️⃣ 審核  
```bash  
python -m packages.rag.generation.Critique\_llm.run --answers outputs/answers.jsonl --out outputs/answers\_pass.jsonl  
```  
  
---  
  
### 5️⃣ RAG / 微調資料輸出  
```bash  
python scripts/export\_rag\_corpus.py outputs/answers\_pass.jsonl  
```  
輸出：  
- `outputs/rag\_corpus.jsonl`  
- `outputs/finetune\_dataset.jsonl`  
  
---  
  
## 🐳 Docker 部署  
```bash  
docker compose up -d --build  
docker compose exec rag-app /entrypoint.sh pipeline  
```  
掛載卷：  
- `/app/data` → 原始資料  
- `/app/indices` → Chunk 索引  
- `/app/outputs` → 生成結果  
  
---  
  
## ✅ 驗收檢查  
- [ ] 問題數量結構正確（每 chunk 11 題）  
- [ ] 答案引用率 ≥ 90%  
- [ ] 審核通過率達標  
- [ ] API / 前端可查詢  
  
---  
  
## 📊 流程圖  
```mermaid  
flowchart TD  
 A[PDF 資料] --> B[Chunk 切分]  
 B --> C[問題生成 + 關鍵字]  
 C --> D[作答]  
 D --> E[審核]  
 E --> F[RAG 資料輸出]  
 E --> G[微調資料輸出]  
```  
  
---  
  
## 💡 Tips  
- \*\*LLM\_MODE\*\*  
 - `MOCK`：離線模式，快速測試流程  
 - `OPENAI\_COMPAT`：對接本地 / 雲端 OpenAI 相容 API  
- \*\*部署\*\*：建議用 Docker 封裝，確保環境一致性  
- \*\*微調\*\*：輸出 `finetune\_dataset.jsonl` 可直接進行 SFT