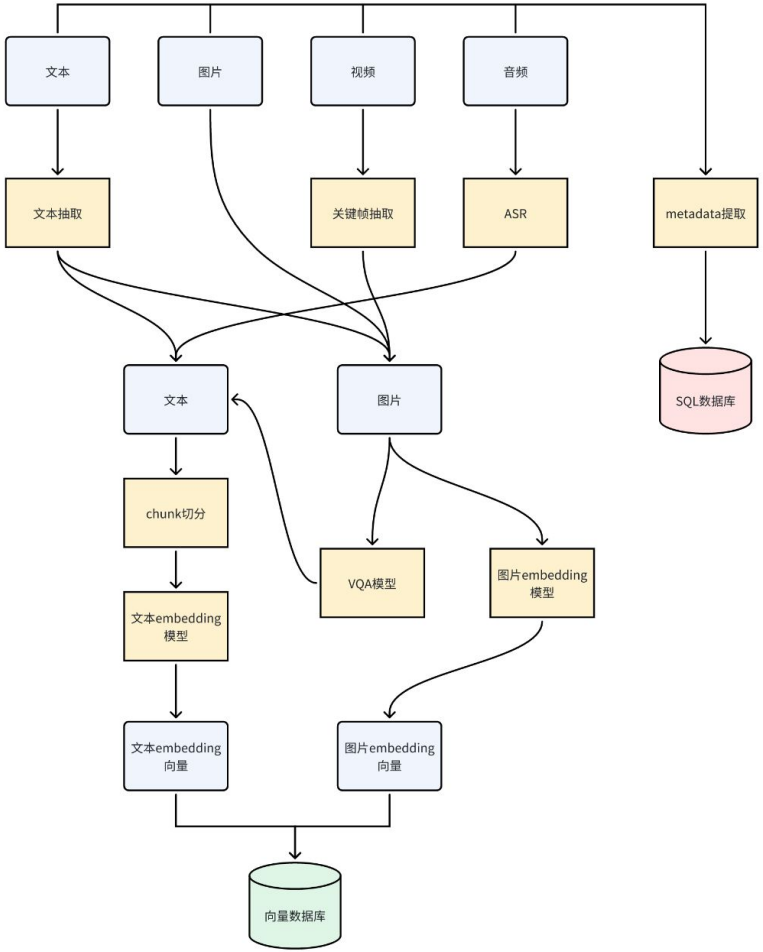
第一次课

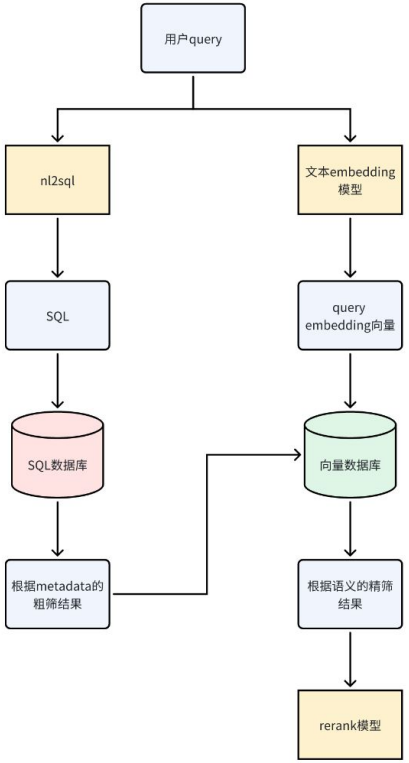
项目1 - 个人知识库

- 构建一个面向用户上传文件的智能处理平台，基于多模态embedding模型与大LLM，实现高效的文件检索、问答交互和智能分类功能
- 技术栈
  - 前后端：Gradio/React.js、FastAPI、Docker、CI/CD、MySQL
  - AI：多模态检索、RAG、向量数据库 (Milvus)、agent、模型微调 (LlamaFactory)、模型推理优化 (vLLM)
  - Cloud：AWS (Lambda, Batch, Sagemaker, S3)

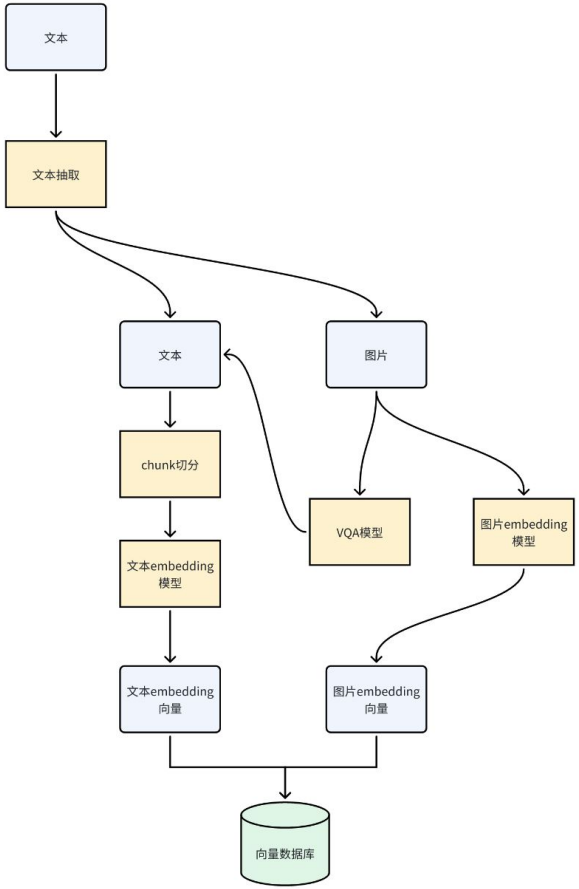
文件输入流程



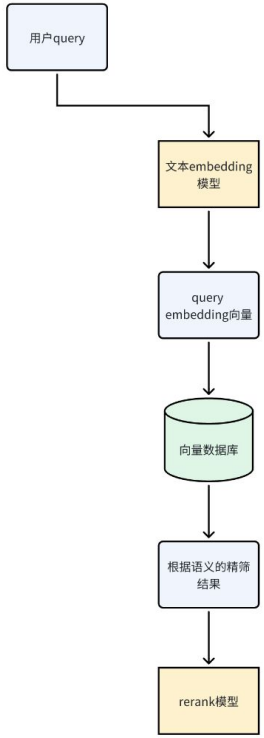
文件检索流程



文件输入流程



文件检索流程



文本抽取

- PyMuPDF (start with this)
  - https://pymupdf.readthedocs.io/en/latest/
- MinerU (recommended)
  - https://github.com/opendatalab/MinerU
  - Tutorial：https://mineru.readthedocs.io/en/latest/user_guide/tutorial/pipeline.html
- markitdown
  - https://github.com/microsoft/markitdown

chunk切分

- https://masteringllm.medium.com/11-chunking-strategies-for-rag-simplified-visualized-df0dbec8e373

文本embedding

- OpenAI API
  - https://platform.openai.com/docs/guides/embeddings
- Open Source
  - https://huggingface.co/spaces/AIR-Bench/leaderboard

图片embedding

- OpenAI API
- https://huggingface.co/spaces/vidore/vidore-leaderboard
- 暂时不要使用Colpali系列模型
- 先使用BGE模型

VQA

- OpenAI API
- Open Source
  - Qwen2.5-Omni
  - Qwen2.5-VL-7B
  - Qwen2.5-VL-3B
  - InternVL3-14B
  - InternVL3-8B
  - Ovis2-8B
  - Ovis2-4B

Weaviate

- 安装:https://medium.com/@newbing/how-to-install-the-weaviate-vector-database-on-debian-12-be5eb121391c
- 数据导入:
  - https://weaviate.io/developers/weaviate/manage-data
- 数据查询:
  - https://weaviate.io/developers/weaviate/search

Milvus

- https://milvus.io/docs

前后端

- FastAPI
  - https://fastapi.tiangolo.com/
- Gradio

示例项目

Todo

- 创建一个github项目
- 看一遍slides提到的技术
- 尝试构建一个纯文本检索系统