Zonda(数字人) -> Imola(RAG Server) -> 查询向量数据库(Qdrant)并将结果(top3)和用户问题一并提交给 llm server(Ollama **mistral:7b-instruct**) -> 返回给Zonda -> tts ->语音转换给数字人说话

Huayra(原始数据导出到向量数据库)

嵌入模型

钻孔偏斜 **BAAI/bge-large-zh**

邻孔间距 **BAAI/bge-large-zh**

终孔偏斜 **BAAI/bge-large-zh**

冻结工程AI大模型文档 **BAAI/bge-large-zh**

智能钻孔静态文档 **BAAI/bge-large-zh**

吉安简介 **BAAI/bge-large-zh**

向量数据库 Qdrant + Qdrant web ui

Langchain来构建其他工具的集成

可提升点：

1. 流式输入和流式输出，大模型输出部分
2. 语音输出更加连贯，数字人部分

银色的手机截图

AI 生成的内容可能不正确。