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。

中间过程结局的关键问题

嵌入器Embedded问题

AnythingLLM自带的嵌入每次都会嵌入过长的向量，很难如自己向向量数据库存储的时候那么精准，再就是存储到qdrant向量数据库的时候，anythingllm没法选择直接去读取已经存好的向量，没法指明哪些表需要，哪些表不需要，因此决定自己去做，完全越过anything llm需要做的事情

大语言模型LLM问题

从deep seek R1切换为更适合处理结构化数据且更适合做本地知识库的模型，**mistral:7b-instruct，从向量数据库中取向量的时候，选择了Top3的方式，嵌入模型选择了BAAI/bge-large-zh**

经过了如上的RAG优化，使当前本地知识库系统可以处理结构化和自然语言类型的文档数据。