**AI LLM assistant project:**

* **SQL database:** store structured personal facts, goals, events (e.g., groceries, workouts).
* **RAG:** retrieve relevant info from SQL using semantic vector search, avoid large context window.
* **LoRA:** fine-tune static user-specific knowledge (name, preferences, long-term goals) into model weights.
* **Shadow chat:** summarize conversations in the back to keep context concise and updated. (both for current talk as well as for storing data)
* **Vector search:** semantic embeddings enable meaning-based retrieval, not keyword matching.
* **Voice mode:** streaming pipeline — streaming ASR (Whisper), fast quantized LLM generation, streaming neural TTS with speaker embedding; async for low latency and natural flow.