Amadeus's Technical Description

Amadeus's Mission

Amadeus is designed to deliver fast, relevant information to wealth managers, enabling them to respond swiftly to client inquiries. By carefully listening to client conversations and synthesizing available data on the fly, Amadeus empowers wealth managers to provide timely and informed answers.

Technological Overview

Transcribing Calls

To accurately capture client requests, Amadeus transcribes calls in real time. For this purpose, it employs the state-of-the-art `Whisper.cpp` model, which offers:

- **High Performance:** Exceptional accuracy even on low-end hardware.
- **Privacy-Friendly Operations:** Capable of on-device transcription to ensure data security.
- **Flexibility:** APIs are available for transcription on alternative hardware setups.
- **Real-Time Processing:** Utilizes a chunking system to transcribe live speech seamlessly.

AI Models for Natural Language Processing

- To extract the semantic meaning of client queries and messages, **WordLlama embedding models** are used due to their speed and lightweight nature. However, any embedding model can be selected based on specific needs. It's important to note that switching models after a certain period requires re-indexing previous data.
- For query understanding, planning, and data explanation, **Gemini 2.0 Flash** is used in the demo. However, with minor prompt adjustments, other high-quality LLMs can be utilized. More privacy-friendly alternatives can also be chosen based on requirements.

Answering Pipeline

Once the calls are transcribed, Amadeus embarks on a multi-step process to generate the optimal response:

1. **Contextual Profiling:** It retrieves the client's profile and portfolio details, integrating this information into the language model's context. This ensures that the responses are tailored to the client's specific needs.
2. **Historical Insights:** Amadeus incorporates relevant data from past client interactions, effectively maintaining long-term memory. The long-term memory is implemented using many bleeding-edge technologies in the RAG space, like Vector Databases, Graph-Based Associations, Hierarchical Clustering, and Memory Decay.
3. **Learning from Experience:** It searches for similar queries previously addressed by wealth managers to continuously refine its response quality.
4. **Strategic Planning:** A call to the language model is made to generate a plan that outlines the key pieces of information required to assist the wealth manager.
5. **Execution:** Finally, the plan is implemented using a suite of tools and search techniques to retrieve the most critical information.