

RAG Pipeline Results

Short Explanation of My Design Choices

For the embedding model, I used **sentence-transformers/all-MiniLM-L6-v2**. I chose this model because it is lightweight, fast, and works well for semantic similarity. It can convert the PDF text into numerical embeddings that allow the system to find information that is related to the meaning of the user's question.

For chunking, I used **sentence-based chunks of 512 tokens with a 50-token overlap**. This keeps each chunk small enough to retrieve specific information while still keeping enough surrounding context. The overlap also helps prevent important information from being separated when the text is split into chunks.

For retrieval, I used a **hybrid retrieval method**, combining vector search and BM25 keyword search. Vector retrieval helps find information based on meaning, while BM25 helps find exact words, technical terms, and numbers. I also used a reranker to place the most relevant retrieved chunks first.

Response to Prompt 1

Prompt: What test methods were used for quality control?

AI-generated response:

The test methods used for quality control include Autoclave for pump tubing, Gamma Irradiation for inlets, Flow Rate Test, Pressure Integrity, Visual Inspection, and Package Integrity.

Response to Prompt 2

Prompt: What are the storage conditions specified in the certificate?

AI-generated response:

The recommended storage temperature is above +5°C. Extended storage below +5°C could lead to brittleness or cracking of the plastic connectors. The operating temperature of the ÄKTA ready flow kits is +2°C to +40°C. The kits should also be allowed to acclimate to a warmer temperature before setup and handling to reduce the risk of damage.