Embeddings & Vector Search – Interview Quick Sheet

1. Load CSV data with Pandas → Clean NaN values.
2. Convert data into dictionary format for processing.
3. Use Sentence Transformers to generate embeddings (all-MiniLM-L6-v2).
4. Store embeddings + metadata in a vector database (Qdrant).
5. Define distance metric (cosine similarity).
6. Insert 1300 records into collection (e.g., 'top_wines').
7. Query using natural language → Encoded into embedding.
8. Retrieve top-k similar results with context.
9. Vector DB enables LLMs to provide relevant answers (RAG pipeline).

Interview Q&A;

Q: Why do we clean NaN values before embeddings?	A: To avoid errors during serialization & embedding.
Q: What role does Sentence Transformers play?	A: It encodes text into numerical embeddings.
Q: Why use cosine similarity?	A: Measures semantic closeness of embeddings.
Q: Why use Qdrant (or any vector DB)?	A: Efficient storage & fast similarity search of embeddings.
Q: How does this connect to LLMs?	A: Embeddings enable LLMs to retrieve context (RAG).

Workflow Visual