Workflow for Round 1B: Persona-Driven Semantic Analysis

This stage uses the highly accurate structural output from the vision-based 1A pipeline to perform a deep semantic analysis across a collection of documents.

Models Used:

- The JSON output from the **Round 1A Vision-Based** workflow.
- A **SentenceTransformer** model (e.g., `all-MiniLM-L6-v2`) for semantic embedding.

Step-by-Step Process:

[Input: Document Collection (3-10 PDFs), Persona, and Job-to-be-Done]

↓ 1. Formulate Semantic Query:

The `persona` and `job-to-be-done` are combined into a single, detailed query string that captures the user's intent.

↓ **2. Content Chunking:** For each PDF, the system uses the JSON output from the Round 1A workflow. It creates logical "chunks" of content by grouping each heading (H1, H2, H3) with its associated paragraph text. ↓ **3. Semantic Embedding:** The **SentenceTransformer** model converts the query string and the content of every chunk into numerical vectors (embeddings) that represent their meaning. ↓ **4. Similarity Calculation & Ranking:** The query embedding is compared against all chunk embeddings using cosine similarity to generate a relevance score for each chunk. All chunks are then ranked together based on this score. ↓ **[Output: Ranked JSON File]** The top-ranked, most relevant chunks are formatted into the final JSON output, including metadata, a ranked list of important sections, and the full text of those sections.