DEMO QUERIES

Dataset:FUNSD
https://www.kaggle.com/datasets/aravindram11/funsdform-understanding-noisy-scanned-documents?resource=download

Test Case 1: Fax Cover Sheet (Noise & Layout)

Filename: `82092117.png` **Document Type:** Confidential Fax Transmission

- **Query 1 (Entity Extraction):**
  "Who sent the fax and what is their phone number?"
  - **Expected Output:** The system retrieves the header block containing the specific sender details.
  - **Target Text:** "SENDER/PHONE NUMBER: June Flynn for Eric Brown/(614) 466-8980"
  - **Why this matters:** Proves the system can extract specific entities (names/numbers) despite OCR noise (e.g., "Brown" vs "Brows").
- **Query 2 (Semantic Understanding):**
  "Is this document confidential? Please quote the disclaimer."
  - **Expected Output:** The system identifies the large warning text box at the bottom.
  - **Target Text:** "THIS MESSAGE IS INTENDED ONLY FOR THE USE OF THE INDIVIDUAL OR ENTITY TO WHOM IT IS ADDRESSED AND MAY CONTAIN INFORMATION THAT IS PRIVILEGED, CONFIDENTIAL..."
  - **Why this matters:** Tests the semantic link between the concept "confidential" and the legal disclaimer paragraph.

Test Case 2: Progress Report (Complex Tables)

Filename: `82200067_0069.png` **Document Type:** Sales Volume Report

- **Query 1 (Table Row Retrieval):**
  "What are the volume and store counts for Dari-Mart?"
  - **Expected Output:** The system retrieves the specific row from the data table.
  - **Target Text:** "Dari-Mart ... 125 / 5 ... 31"
  - **Why this matters:** Proves the embedding model maintains row-level relationships in tabular data, rather than treating words as a "bag of words."

- **Query 2 (Header Information):**
  "Who is the report from and what is the subject?"
    - **Expected Output:** The system retrieves the header section.
    - **Target Text:** "FROM: T. D. Blachly ... SUBJECT: OLD GOLD MENTHOL LIGHTS & ULTRA LIGHTS 100'S - PROGRESS REPORT"

Test Case 3: Product Introduction Form (Handwriting & Layout)

Filename: `82250337_0338.png` **Document Type:** Competitive Product Introduction

- **Query 1 (Handwritten Field Extraction):**
  "Where is the test market geography located?"
    - **Expected Output:** The system retrieves the field where "Wisconsin" is written.
    - **Target Text:** "TEST MARKET GEOGRAPHY: Divisions 621 and 627 (Wisconsin)"
    - **Why this matters:** Demonstrates the OCR's ability to handle handwritten annotations combined with printed text.
- **Query 2 (Descriptive Analysis):**
  "How has the sales force been involved with the distribution?"
    - **Expected Output:** The system retrieves the narrative paragraph.
    - **Target Text:** "They have crew-worked distribution... Sales force has been busy promoting old style packs to clean up inventory."

Test Case 4: Retail Progress Report (Checkbox & Context)

Filename: `82251504.png` **Document Type:** Retail Excel Progress Report

- **Query 1 (Checkbox Logic):**
  "For which month is this submission?"
    - **Expected Output:** The system finds the line marked with an "X".
    - **Target Text:** "October 31 (X)"
    - **Why this matters:** Shows the ability to capture "checked" states in forms, which is critical for processing administrative documents.
- **Query 2 (Sentiment/Feedback Analysis):**
  "Was the Flex Payment program successful with chains?"
    - **Expected Output:** The system retrieves the specific feedback paragraph.
    - **Target Text:** "Chains: This program has been successful to date with chains where our 'Flex Payment' was not in place... where we were using the 'Flex Payment' system we have not been as successful."

Test Case 5: Regional Status Report (Sparse Data)

Filename: `82252956_2958.png` **Document Type:** Progress Report (Variant)

- **Query 1 (Specific Metric Retrieval):**
  "How many stores does Walgreen Drug have?"
  - **Expected Output:** The system retrieves the single data row available.
  - **Target Text:** "Walgreen Drug ... 144/14 ... 93"
  - **Why this matters:** Tests precision in documents that are mostly empty whitespace with isolated data points.
- **Query 2 (Submission Date Confirmation):**
  "What is the submission date marked on the form?"
  - **Expected Output:** The system identifies the checked date box.
  - **Target Text:** "JUN 23 [X]"