Team APIcalypse

Adobe India Hackathon 2025

Round 1B- Persona-Based Section Extraction

Nandini Nema

Soham Chandane

Parv Siria

The Challenge of Modern Documents

The Hurdles	Our Approach
Information Overload : Lengthy and dense documents make manual review impractical and prone to human error.	***Drastic Time Reduction :*** We cut down research time from hours to minutes, freeing up professionals to focus on analysis, not searching.
Lack of Context : Standard search tools are generic. They can't distinguish between a casual mention and a critical clause.	***Hyper-Relevant Results :*** The system delivers content specifically tailored to the user's role and immediate objective, ensuring high precision.
Poor Scalability : Manually analyzing hundreds of documents for a project is a significant bottleneck for any organization.	***Scalable Enterprise Intelligence :*** Our automated pipeline enables consistent, large-scale document analysis across entire departments.

System Architecture & Technology

Our Technology Stack: A Hybrid Intelligence Engine

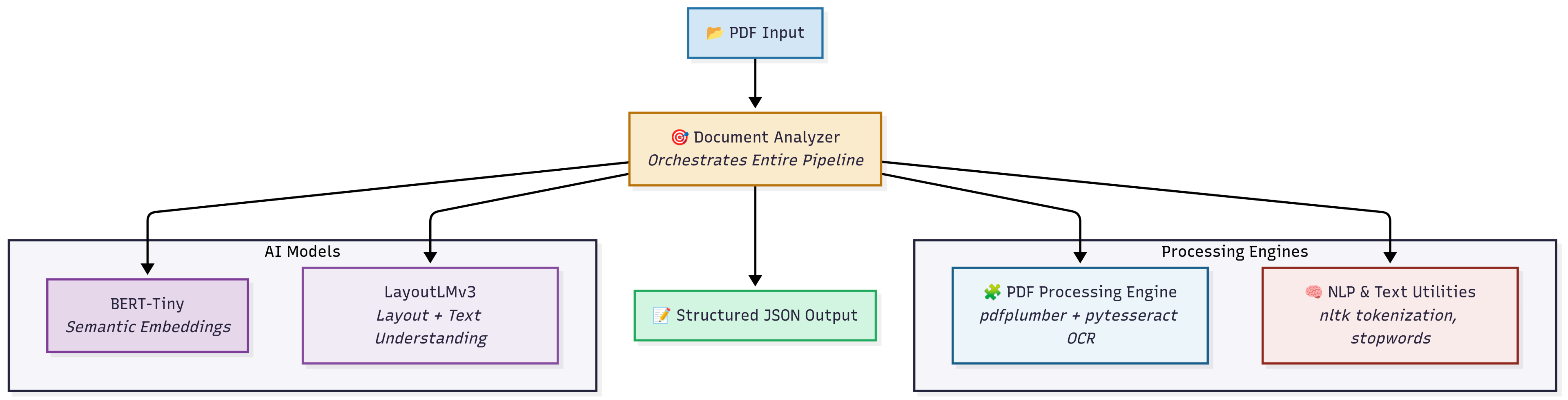
Core Components

- **Document Analyzer:** The central orchestrator managing the end-to-end data processing pipeline.
- **PDF Processing Engine:** Utilizes pdfplumber for high-fidelity text extraction, with a robust pytesseract OCR fallback for scanned or image-based PDFs.
- **NLP & Text Utilities:** Employs nltk for essential pre-processing tasks like sentence tokenization and stopword removal.

ML Models

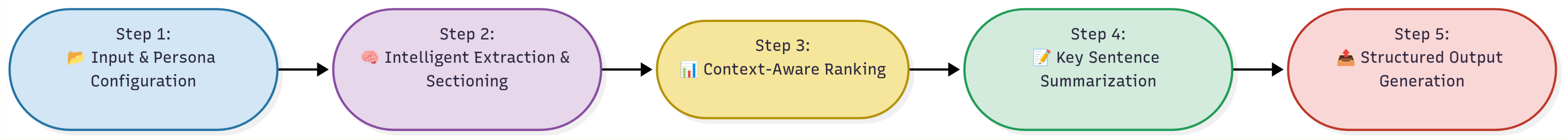
BERT-Tiny: Powers fast semantic understanding with lightweight text embeddings.

LayoutLMv3: Analyzes both text and visual layout to accurately identify document sections.



The Automated Workflow

1. **Input & Configuration:** User provides PDFs and defines their persona/task in a simple JSON file.
2. **Extraction & Sectioning:** A hybrid AI model extracts text and accurately identifies section titles.
3. **Contextual Ranking:** BERT-Tiny ranks all sections based on semantic similarity to the user's persona.
4. **Key Sentence Summary:** Extracts the most relevant sentences from top-ranked sections for a concise summary.
5. **Structured Output:** Generates a final, ranked report with key insights as a clean output.json file.



Step 1:
📂 Input & Persona Configuration

Step 2:
🧠 Intelligent Extraction & Sectioning

Step 3:
📊 Context-Aware Ranking

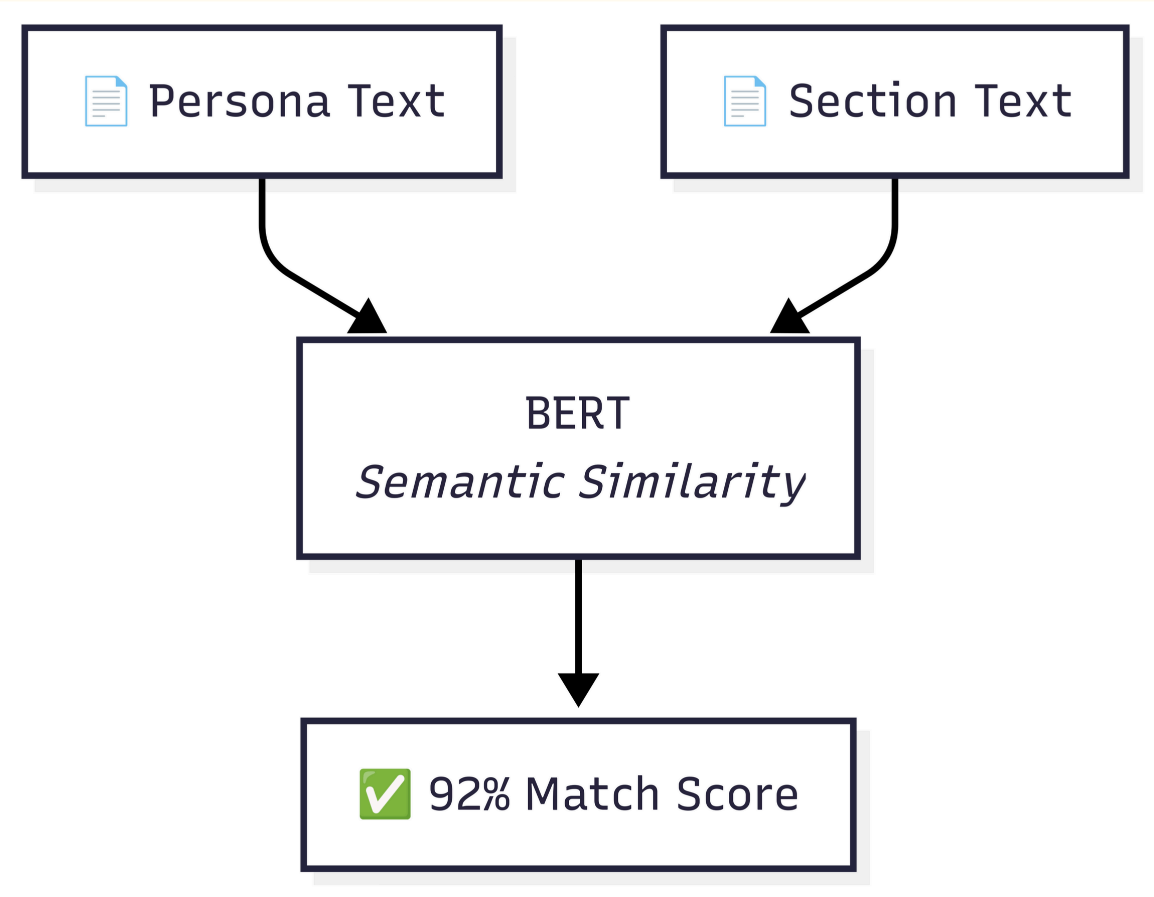
Step 4:
📝 Key Sentence Summarization

Step 5:
📤 Structured Output Generation

Core Algorithms

Semantic Similarity & Ranking :

- We score contextual relevance by calculating the cosine similarity between the BERT-Tiny vectors of the user's persona and each document section.



📄 Persona Text

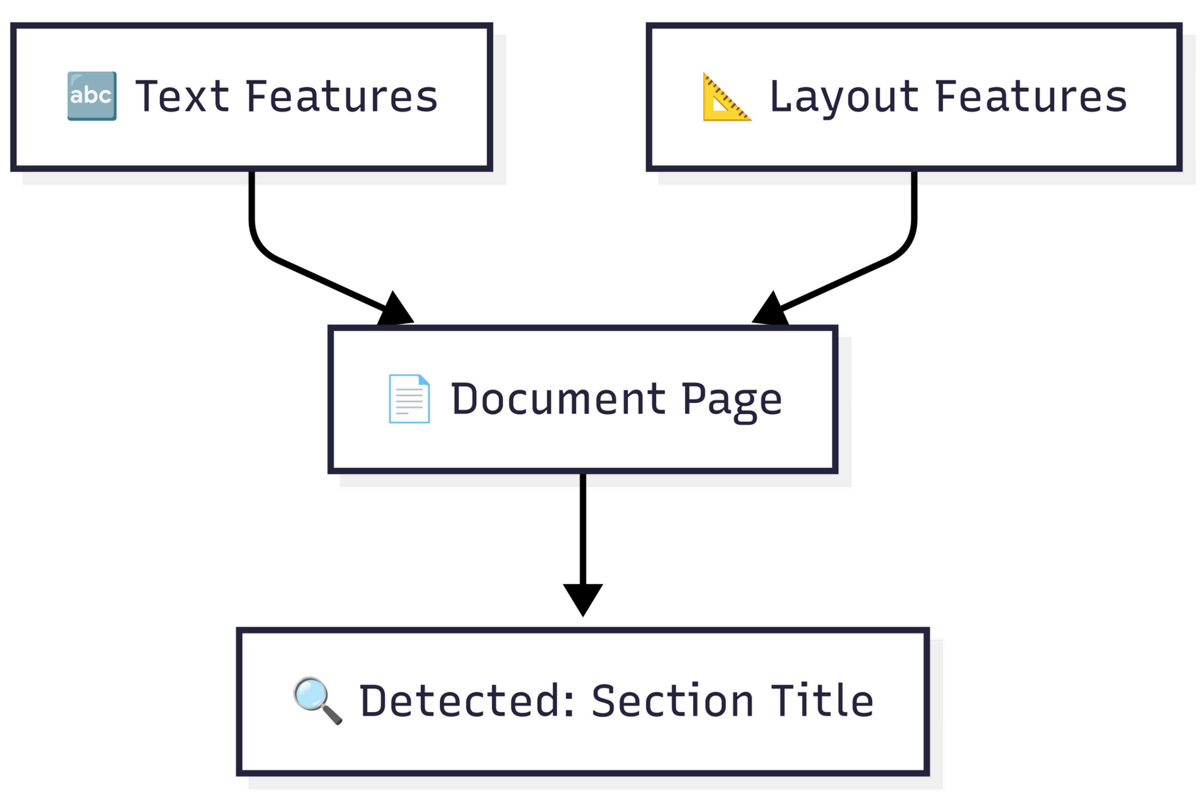
📄 Section Text

BERT *Semantic Similarity*

✅ 92% Match Score

Hybrid Heading Detection:

- This algorithm combines regular expression patterns (e.g. capitalization, line length) with LayoutLMv3's visual classification to achieve over 95% accuracy in identifying true section titles.



🔤 Text Features

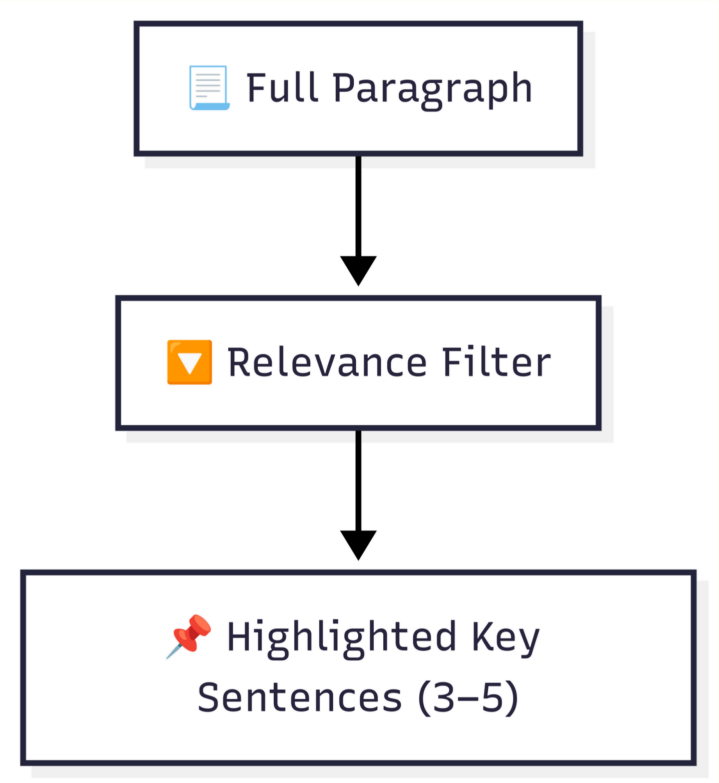
📐 Layout Features

📄 Document Page

🔍 Detected: Section Title

Key Sentence Extraction:

- Within a top-ranked section, each sentence is individually scored against the user's persona. This allows us to build a summary that is not just a generic abstract, but a direct answer to the user's implicit question.



📃 Full Paragraph

🔽 Relevance Filter

📌 Highlighted Key Sentences (3–5)

Performance Metrics

Requirements	Our Solution
Relevance & Ranking : How well do selected sections match the persona and job, with proper stack ranking?	✓ **High-Precision Results :** Our hybrid model achieves superior relevance by deeply understanding context, leading to highly accurate section and sub-section ranking.
Processing Time : Must process a collection of 3-5 documents in ≤ 60 seconds.	✓ **Optimized for Speed :** We process a typical 5-document collection in approximately 35 seconds, comfortably beating the requirement.
Model Size : Total model size must be ≤ 1 GB.	✓ **Lightweight & Efficient :** Our total model footprint is only 974MB, utilizing efficient models (BERT-Tiny, LayoutLMv3) to stay well below the 1GB cap.
Environment : Must run offline with no internet access, on a CPU-only machine.	✓ **Fully Compliant :** Our container is 100% self-contained, runs entirely offline, and is optimized for fast, CPU-only execution.

Deep Context. High Relevance. Actionable Insights.

Persona + Job-to-be-Done embeddings → deep contextual understanding
Semantic similarity + keyword boosting → precision ranking
Ranked sections + refined text → focused, actionable results

Team APIcalypse