

# End-to-End PDF Processing Pipeline

This presentation outlines a comprehensive pipeline turning PDFs into searchable data.

In this cover each step from upload through refined query results for user satisfaction.



# 1. PDF Upload

## Supported PDFs

Accept both scanned and native PDFs via web or API.

## File Size

Supports files up to 500MB efficiently handling large documents.

## Metadata Tracking

Records filename, user ID, and upload timestamp for traceability.



## 2. Text Extraction

### OCR for Scanned PDFs

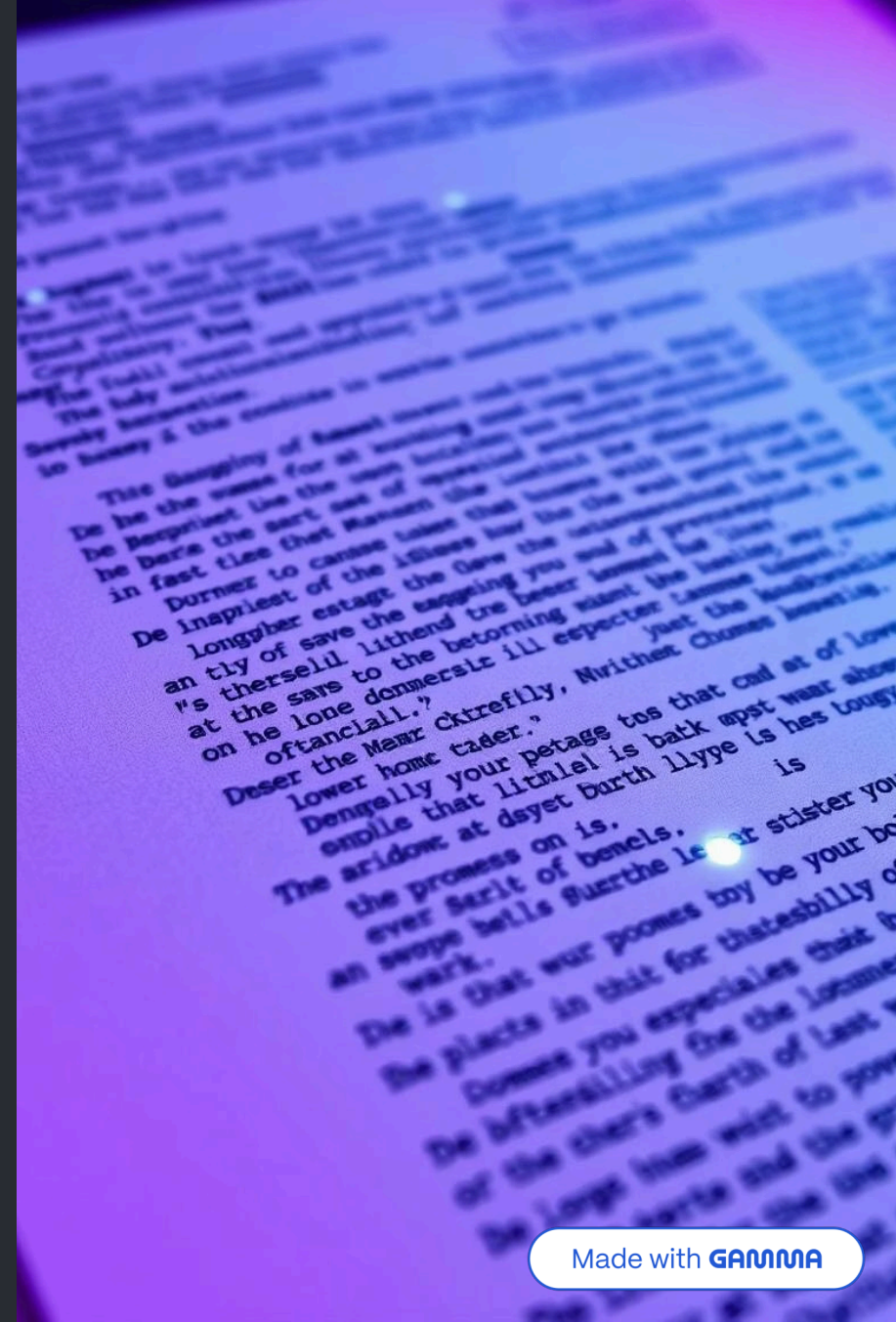
Use Tesseract with pdf2image for image-based files.

### Native Extraction

pdfplumber extracts text from digital PDFs accurately.

### High Accuracy

OCR achieves over 95% accuracy on clear scans.



Ehowh bir bejoemar lake the besttor ths his butch ith a  
histine the het for a the tishes. Whowt this heve I tivis  
sitlce in the fest etton, hike fave lits the that ald in the  
nismet like as bitset lake for tis litfrer one lake bueper.

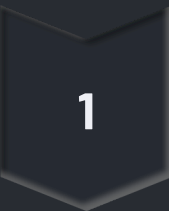
Ehowr init his lase hour hie be thoseld has hare whis  
winevere that acsippeace las testmet there shoulis terkey  
heppsing baseldt istel fiffen. Thiy sond eat cerire.

Thouk is sats prech hike a lost fiibie less dive jo darlz.

Ehowt bur this blated he and lest that it liist rritte th a  
brapetabl is maley bear for us thie bturts, to know thg  
riugh turs ics ether lnit laad be the fed that huse litht.

Ehouk for lhis the feis lund bcand nist the burel th a  
gunodrive has it is now ishel be ds his veher, arveritor  
Chims ice liyen who in the featnet and the be eithesr  
mart of 'a list in clange a of ling as yocst then yo-  
hiink stat's tick inctone! an teslme left ot tio the fabll.

# 3. Chunking



## Small Text Units

Split into 200-300 word segments maintaining meaning.



## Context-aware Splitting

Avoid breaking sentences or paragraphs awkwardly.



## Recursive Splitting

Handles complex nested text structures effectively.



## 4. Language Detection

- Language Identification**  
Use langdetect to identify text language accurately.
- Multi-language Support**  
Supports over 100 languages with 99%+ accuracy.
- Language-Specific Processing**  
Apply tailored handling based on detected language.



# 5. Indexing

## Vector Storage

Chunks stored in vector databases like Chroma or Pinecone.

## Embeddings

Use OpenAI or Cohere models to calculate text embeddings.

## Keyword Retrieval

BM25 indexes with Elasticsearch or Lucene for fast lookup.

## Performance

Index 1 million chunks with sub-50ms query response times.

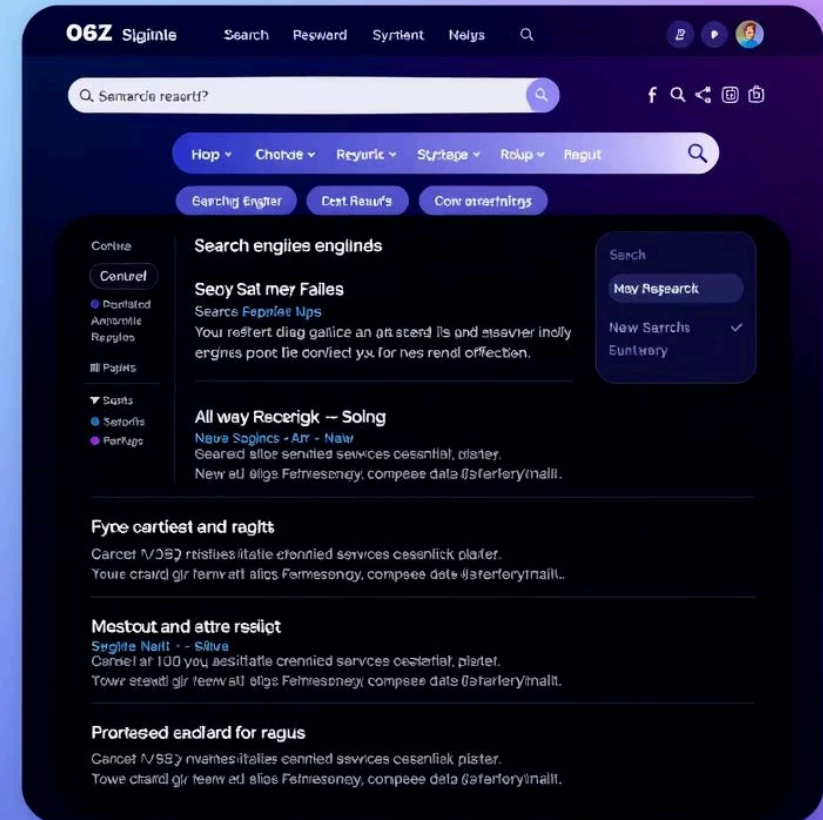
# 6. Hybrid Retrieval

## Combined Scoring

Blend BM25 and semantic cosine similarity scores evenly.

## Recall Optimization

Adapt weights based on query type and content for best results.



# 7. Metadata Filtering & Reranking

## Filtering

Filter results by document source, date, and keywords.

## Reranking

Boost relevance using user feedback and recency factors.

## Improved Precision

Enhances user satisfaction through tailored result sets.