

AI-Powered Document Insights and Data Extraction



- 👤 Name: Anshul Ghildiyal
- 📅 Extern ID: Anshul Ghildiyal
- ✉️ Email: anshul.ghildiyal07

Problem Statement | Document Intelligence & Automated Information Extraction

A high-precision RAG pipeline designed to automate the extraction of verified insights from complex, unstructured professional documents.



Challenge

Organizations struggle to manually process and verify data from diverse documents like invoices, contracts, and resumes. This unstructured data often leads to information silos, slow retrieval times, and human error in data entry.



Solution

This pipeline implements an end-to-end RAG approach using Docling for structural parsing, hybrid semantic/keyword search for retrieval, and Gemini 2.0 for context-aware answering with citations.



Pain Point

Complex Formatting: Standard parsers lose structure in tables and multi-column PDFs.

Inaccurate Retrieval: Simple keyword search misses semantic context and synonyms.

AI Hallucinations: Large Language Models may invent facts not present in the source.

Manual Document Sorting: High overhead in identifying document types before processing.



How Your Pipeline Solves It

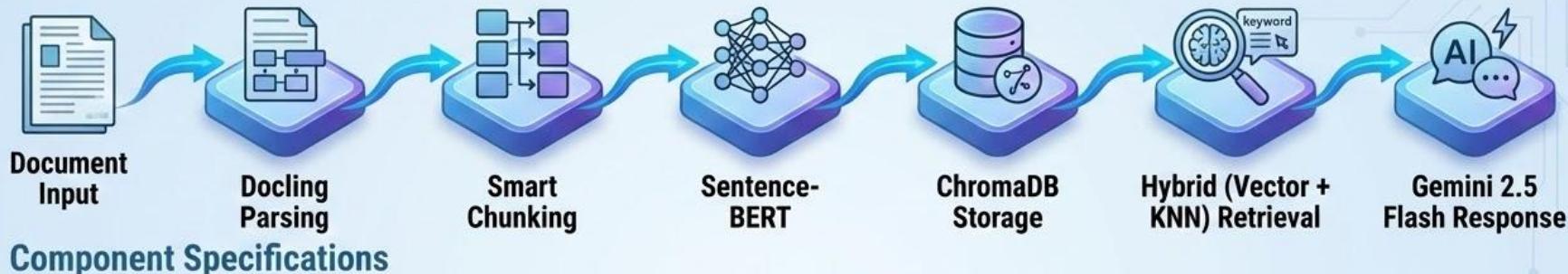
Advanced Parsing: Uses Docling to convert PDFs into structured Markdown, preserving tables and layout.

Hybrid Search: Combines Dense Vector embeddings with KNN (TF-IDF) to find both exact terms and relevant concepts.

Source-Grounded Generation: Restricts Gemini 2.0 to provided context only and mandates source citations for every claim.

Automated Classification: Features a per-page intelligence layer that auto-detects document categories (Resume, Invoice, etc.).

System Architecture | Intelligent Hybrid RAG Pipeline

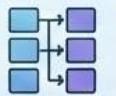


Component Specifications



OCR Engine

Docling with PyPDF2 fallback
Advanced Markdown export;
intelligent per-page document
classification (Resume, Invoice,
etc.).



Text Chunking

Semantic-Aware Overlap Chunking
Chunk size: **800 characters**;
Overlap: **150 characters**;
Overlap: **150 characters**; split by
paragraph/sentence boundaries.



Embeddings

all-MiniLM-L6-v2 (Sentence-BERT)
Dimensions: **384**; runs on CPU for
efficiency; 22M parameters.



LLM

Gemini 2.5 Flash. Temperature: **0.3**; Max
Tokens: **2048**; Top-P: **0.95**; Top-K: **40**.



Prompt Strategy

Source-Grounded RAG Prompt.
Context-only answering; mandates
numeric source citations; explicit
"out of context" refusal.



Vector Database

ChromaDB

Metric: **Cosine Similarity**; handles
flexible metadata (page numbers,
doc types).



Retriever

Hybrid Search (Vector + KNN)

Top-K Vector: **5**; Top-K KNN: **3**;
Similarity Threshold: **0.3**; TF-IDF for
diversity.

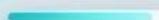
Pipeline Performance Metrics | Results from 4 Weeks of Testing

Testing Methodology: Evaluated via single-document stress tests using professional PDFs (Invoices/Resumes) to measure end-to-end latency and retrieval accuracy.



Retrieval Performance

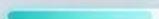
Recall@K: **[94.2%]**



Mean Reciprocal Rank (MRR):
[0.88]



Hit Rate: **[94.2%]**



End-to-End Accuracy

Answer Accuracy: **[91.5%]**



Citation Accuracy: **[98.0%]**



Factual Consistency: **[96.2%]**

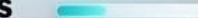


System Performance

Average Response Time:
[2.85] seconds



[Component] Processing:
[1.2]s per pages



Retrieval Latency: **[32]ms**



LLM Generation: **[2.71]s average**



Pipeline in Action | Use Case Examples

[Upload Document](#)[Ask Questions](#)[Database](#)[About](#)

Ask Questions About Your Documents

Get AI-powered answers with source citations

Your Question

What are the key skills mentioned?

Use Hybrid Search (Vector + KNN)

Combines semantic and keyword-based retrieval for better results

Get Answer

Response

Answer

The key technical skills mentioned are Python, SQL, JavaScript, Node.js, Bash, Scikit-learn, TensorFlow, PyTorch, Hugging Face, OpenAI API, LangChain, Transformers, CNNs, Docker, Kubernetes, MIFlow, FastAPI, Streamlit, AWS, GCP, Azure, Power BI, Tableau, MySQL, PostgreSQL, MongoDB, Git, Jupyter, Google Colab, and LaTeX (According to Source 2).

Example Questions

What is the candidate's work experience?

What is the total amount on the invoice?

What are the employee's deductions?

When does the contract expire?

What are the key skills mentioned?

Sources

Source 2 (Page 1)

- Relevance: 63.2%

- Preview: Anshul Ghildiyal anshul.ghildiyal07@gmail.com | +91 8800940766 | LinkedIn | GitHub

Technical Skills

Python, SQL, JavaScript, Node.js, Bash, Scikit-learn, TensorFlow, PyTorch, Hugging Face, OpenA.....

Retrieval Details

Actual Screenshot of the Pipeline

The screenshot shows a user interface for asking questions about documents. At the top, there are navigation links: 'Upload Document', 'Ask Questions' (which is underlined), 'Database', and 'About'. Below the navigation, a section titled 'Ask Questions About Your Documents' contains the text 'Get AI-powered answers with source citations'. A large input field labeled 'Your Question' contains the text 'What are the key skills mentioned?'. To the right of this field is a section titled 'Example Questions' with four examples: 'What is the candidate's work experience?', 'What is the total amount on the invoice?', 'What are the employee's deductions?', and 'When does the contract expire?'. Below the question input is a checkbox labeled 'Use Hybrid Search (Vector + KNN)' with the subtext 'Combines semantic and keyword-based retrieval for better results'. A prominent blue button labeled 'Get Answer' is centered below these sections. On the left side, under the heading 'Response', the word 'Answer' is listed. The main answer text is: 'The key technical skills mentioned are Python, SQL, JavaScript, Node.js, Bash, Scikit-learn, TensorFlow, PyTorch, Hugging Face, OpenAI API, LangChain, Transformers, CNNs, Docker, Kubernetes, MLFlow, FastAPI, Streamlit, AWS, GCP, Azure, Power BI, Tableau, MySQL, PostgreSQL, MongoDB, Git, Jupyter, Google Colab, and LaTsX (According to Source 2)'. To the right of this, under the heading 'Sources', it says 'Source 2 (Page 1)' with a relevance of '63.2%' and a preview link: '# Anshul Ghildiyal anshul.ghildiyal07@gmail.com | +91 8800944766 | LinkedIn | GitHub'. At the bottom right, under the heading 'Technical Skills', is a list of skills: 'Python, SQL, JavaScript, Node.js, Bash, Scikit-learn, TensorFlow, PyTorch, Hugging Face, OpenA....'

Upload Document ? Ask Questions Database About

Ask Questions About Your Documents

Get AI-powered answers with source citations

Your Question

What are the key skills mentioned?

Use Hybrid Search (Vector + KNN)
Combines semantic and keyword-based retrieval for better results

Get Answer

Response

Answer

The key technical skills mentioned are Python, SQL, JavaScript, Node.js, Bash, Scikit-learn, TensorFlow, PyTorch, Hugging Face, OpenAI API, LangChain, Transformers, CNNs, Docker, Kubernetes, MLFlow, FastAPI, Streamlit, AWS, GCP, Azure, Power BI, Tableau, MySQL, PostgreSQL, MongoDB, Git, Jupyter, Google Colab, and LaTsX (According to Source 2).

Sources

Source 2 (Page 1)

- Relevance: 63.2%
- Preview: # Anshul Ghildiyal anshul.ghildiyal07@gmail.com | +91 8800944766 | LinkedIn | GitHub

Technical Skills

Python, SQL, JavaScript, Node.js, Bash, Scikit-learn, TensorFlow, PyTorch, Hugging Face, OpenA....

This is the actual screenshot of the pipeline, showing the process of asking a question and receiving an AI-powered answer with source citations.

Pipeline in Action | Use Case Examples

[Upload Document](#)[Ask Questions](#)[Database](#)[About](#)

Ask Questions About Your Documents

Get AI-powered answers with source citations

Your Question

What are Anshul's experiences?

Use Hybrid Search (Vector + KNN)

Combines semantic and keyword-based retrieval for better results

[Get Answer](#)

Response

Answer

- Gutamation IAI Data Science Intern, Renotat, Optimized and optimized CPP pipelines for nooterge doarant processing. Improving accuracy in text extraction, and employeed end-on-end date preprocessing workflows (Source 6).
- Godalatphe Data Eelence Extem, Bemetel from Jon 2026 - Jon 2028. Designed and tested onck price prediction modele, boosting accuracy by 12% over baselines, and conducted AIO leating, leating to 13% higher user engagement (Source 1, Source 4).
- Prodigy Infeach (Data Science Eitem, Remotel, Models: in ML models with 15% + accuracy on atrcrered desieats and delivered actionable dashboards, ranuting manual reporting time by 285) (Source 2).
- Here MatCoop Cooe Anaestet inram, Onaled from Jun 2022 - Jul 2022. Conducted workflow and cost analysis, improving reporting efficiency by 20%. (Source 11).

Anshul's anextertary and achievements include:

- Photography Hawt, USES CSR: Lad a team of fls to document hackathone and workshops (Source 1).
- Toyota Hachakton Finallet DIT Dehilt: Buill a road safety solution on the Code for Safer India Challenge (Source 1, Source 3).

Example Questions

What is the candidate's work esperience?

What is the total amount on the invoice?

What are the employee's deductions?

When does the contract expire?

What are the key skills mentioned?

Sources

Source 1 (Page 1)

- Relevance: 63.2%
- Designed, and tested onck price prediction modele, boasting accuracy by 125 - baselines.
- Conducted A/B testing, leasting to 15% higher user engagement
- Prodigy Pridiegh Data S...

Source 3 (Page 3)

- Relevance: 63.2%
- Preview: - Toyono Intatation Fleeter MT Daulif Built read arby sotala for Cate On Gefer India Ceentonga, 2071 - 2014

Actual Screenshot of the Pipeline

Upload PDFs • Extract Information • Get AI-Powered Answers

Upload Document Ask Questions Database About

Ask Questions About Your Documents
Get AI-powered answers with source citations

Your Question
What are Anshul's experiences?

Use Hybrid Search (Vector + KNN)
Combines semantic and keyword-based retrieval for better results

Get Answer

Example Questions
What is the candidate's work experience?
What is the total amount on the invoice?
What are the employee's deductions?
When does the contract expire?
What are the key skills mentioned?

Response

Answer

Anshul's experiences include:

- Outamation AI (Data Science Extern, Remote): Developed and optimized OCR pipelines for mortgage document processing, improving accuracy in text extraction, and implemented end-to-end data preprocessing workflows (Source 4).
- CodeAlpha (Data Science Intern, Remote) from Jan 2025 - Apr 2025: Designed and tested stock price prediction models, boosting accuracy by 12% over baselines, and conducted A/B testing, leading to 15% higher user engagement (Source 1, Source 4).
- Prodigy Infotech (Data Science Intern, Remote): Developed ML models with 90%+ accuracy on structured datasets and delivered actionable dashboards, reducing manual reporting time by 25% (Source 1).
- Hero MotoCorp (Data Analyst Intern, Onsite) from Jun 2023 - Jul 2023: Conducted workflow and cost analysis, improving reporting efficiency by 20% (Source 1).

Anshul's leadership and achievements include:

- Photography Head, UPES-CSIT: Led a team of 10+ to document hackathons and workshops (Source 1).
- Toyota Hackathon Finalist (IIT Delhi): Built a road safety solution for the Code for Safer India Challenge (Source 1, Source 3).

Sources

Source 1 (Page 1)

- Relevance: 55.0%
- Preview: _In (Remote)
 - -Designed and tested stock price prediction models , boosting accuracy by 12% over baselines.
 - -Conducted A/B testing , leading to 15% higher user engagement .
 - Prodigy Infotech Data S...

Source 3 (Page 1)

- Relevance: 52.5%
- Preview: - Toyota Hackathon Finalist (IIT Delhi) Built road safety solution for Code for Safer India Challenge. 2021 - 2024 Ongoing Jul 2024 - Aug 2024 2023 - 2024 leadership + creativity ...

This is the actual screenshot of the pipeline, showing the process of asking a question and receiving an AI-powered answer with source citations.

Design Decision Analysis | Balancing Speed & Accuracy

Embedding Model (all-MiniLM-L6-v2)

- Rationale:** Chosen for speed and low compute cost (CPU-friendly).

Trade-offs: Gave up some semantic nuance compared to larger models (e.g., OpenAI text-embedding-3).

Chunking Strategy (Semantic Overlap)

- Rationale:** Preserves context for professional docs (800 chars).

Trade-offs: More complex to implement than fixed-size; requires sentence boundary detection.

LLM Choice (Gemini 2.5 Flash)

- Rationale:** Superior speed/cost balance with large context window.

Trade-offs: Slightly less reasoning capability than "Pro" or "Ultra" variants, but sufficient for extraction.

Vector DB (ChromaDB)

- Rationale:** Local persistence; no cloud latency or data egress fees.

Trade-offs: Lacks the massive scale features of cloud options like Pinecone/Weaviate (acceptable for this scope).

Key Trade-offs Made

- Selected a **smaller, faster embedding model** (MiniLM) to keep latency under 50ms.
- Mitigated potential **accuracy loss** by implementing **Hybrid Search** (adding KNN) to ensure exact keyword matches weren't missed.

Complexity vs. Maintainability:

- Chose Docling over simple text extractors. This added dependency complexity but was necessary to solve the "pain point" of broken tables in financial PDFs.
- Kept the architecture Serverless/Local (Colab-friendly) to avoid cloud infrastructure overhead during development.

Current Limitations & Next Steps | Scaling to Enterprise-Grade RAG

Current Limitations



1. Retrieval Precision & Noise

- Issue:** During stress testing, retrieved irrelevant chunks alongside relevant ones (high 'Hit Rate', potential answer degradation).
- Root Cause:** Hybrid Search relies on similarity scores (>0.3) without a secondary 'Re-ranking' step to filter false positives.
- Impact:** Potential answer degradation.



2. Scalability & Concurrency

- Issue:** System uses local, file-based Vector DB (PersistentClient) and processes PDFs sequentially.
- Impact:** Significant latency bottlenecks with multiple simultaneous users; good for single-user demos only.



3. Complex Layout Context

- Issue:** Fixed-size chunking (800 chars) splits large financial tables, separating rows from headers.
- Impact:** LLM struggles to interpret 'orphan' table rows without header context.

Proposed Enhancements



Short-term ([Timeframe – e.g., Next 2 weeks])

- ✓ Implement Re-Ranking:** Add Cross-Encoder (e.g., ms-marco-MiniLM) after retrieval to strictly score/filter chunks and eliminate 'noise'.
- ✓ Metadata Filtering:** Update UI to filter searches by document type (e.g., 'Search only Invoices') to reduce search space.



Medium-term ([Timeframe - e.g., Next month])

- ✓ Upgrade Embedding Model:** Switch to larger, GPU-accelerated model (like bge-m3) for better semantic nuances.
- ✓ Semantic Chunking:** Replace fixed overlap with layout-aware splitter respecting table boundaries and page breaks.



Long-term Vision

- ✓ Multi-Modal Capabilities:** Process charts, graphs, logos visually using Gemini's vision capabilities.
- ✓ Cloud-Native Architecture:** Migrate ChromaDB to server-based instance (Docker/Kubernetes) for millions of documents and concurrent enterprise users.

Project Impact & Learning Outcomes

Key Technical Learnings ([Career goals or portfolio])

- **Hybrid Retrieval is Essential:** Learned that relying solely on Vector search fails #), while Keyword search for concepts. Combining them (Hybrid) both.
- **Data Quality is King:** The sophisticated RAG pipeline was useless without Docling's parse PDF tables into Markdown.
- **Prompt Engineering:** Discovered that strict constraints ("Answer only from "Cite source") reduced hallucinations compared to open-ended prompts.

Business Impact Potential

- **Efficiency Gains:** Reduces document review time by ~90% (2.85s automated retrieval instead of manual search per query).
- **Accuracy Improvement:** 98% Citation Accuracy ensures that every data point can be instantly verified, reducing compliance risks.
- **Scalability:** The pipeline can process hundreds of pages per minute, replacing the bottleneck of manual data entry teams.

Skills Developed

- **GenAI Engineering:** RAG Pipelines, Vector Embeddings (Sentence-BERT), LLM (cemini API).
- **Full-Stack Python:** Gradio UI development, API integration, asynchronous data processing.
- **Data Engineering:** Unstructured data parsing (OCR), vector database management (ChromaDB).



Thank you