# 📘 Metadata-Enriched PDF Ingestion for AI Applications

## 🧾 Objective

This initiative establishes a comprehensive ingestion pipeline designed to process PDF documents, enrich them with metadata, generate semantic embeddings using Gemini, and index them into AI applications for precise semantic search and retrieval-augmented generation (RAG). The metadata facilitates enhanced traceability and accuracy for document question-answering tasks.

## 🔧 Tools & Technologies

| Component | Role |
| --- | --- |
| Google Cloud Storage | Storage for raw PDF files |
| Vertex AI (Gemini) | Generation of semantic embeddings |
| AI Applications | Indexed storage for semantic retrieval |
| pdfplumber | Structured text extraction with layout awareness |
| Python SDKs | Interaction with GCS, Vertex AI, and AI Applications |

## 🔍 Metadata Strategy

Each chunk generated from PDFs is enriched with comprehensive metadata to improve semantic searchability and context preservation. Metadata includes:

| Field | Description |
| --- | --- |
| source\_file | Original PDF file name |
| page | Page number from which text chunk was extracted |
| paragraph\_index | Position of paragraph within the page |
| section\_heading | Section or heading title (optional, when available) |

This metadata is explicitly stored within AI applications to facilitate easy and accurate retrieval.

## 🧩 Chunking Strategies

The pipeline leverages two primary chunking strategies:

### Paragraph-based Chunking:

* PDF documents are parsed into paragraph-based chunks.
* Metadata such as page numbers, paragraph indexes, and section headings are included.

### Token-based Chunking:

* PDFs are divided into chunks of approximately 500 tokens with a 125-token overlap.
* This strategy ensures contextual continuity across chunk boundaries.
* Token-based chunking supplements paragraph-based chunking for flexible retrieval scenarios.

## ⚙️ Implementation Workflow

### Environment Configuration

* Authentication via Google Colab and service accounts.
* Setting environment variables for Google Cloud project, region, and storage buckets.

### PDF Processing

* Files stored in Google Cloud Storage are listed and downloaded for local processing.
* Text extraction performed using pdfplumber.

### Chunk Generation

* Implement custom functions for both paragraph-based and token-based chunking.
* Enrich chunks with metadata.

### Embedding Generation

* Semantic embeddings generated using Vertex AI Gemini embedding model.

### Data Indexing

* Chunks with embeddings and metadata are indexed into AI applications.
* Structured metadata ensures enhanced searchability and contextual retrieval.

### Verification & Testing

* Confirm successful ingestion and indexing by performing semantic searches.
* Validate metadata retrieval accuracy such as file names and page numbers.

## ✅ Outcome & Validation

* Successfully stored metadata-enriched chunks.
* Accurate retrieval and display of metadata in query results.
* Comprehensive testing confirms the effectiveness of metadata integration in AI-driven retrieval applications.

## 📌 Next Steps

* Further refine section heading detection.
* Enhance metadata with additional contextual fields.
* Monitor and iterate based on retrieval performance metrics.