

Intelligent Document Processing System

Automate, Validate, Translate, Summarize for
Business-Critical Documents Seamlessly

Overview

Our Platform leverages Google/Azure Document AI for Extraction, enabling a accurate, Schema Driven Data extraction, Human in the loop Validatio and Multilingual translation. Leverage LLM Models eg (GeminiPro) for Intelligence capability

Features

Document and Schema Upload

Easily upload documents (PDF,Images) and Define extraction schemas

Confidence Scoring

Visualizing extraction accuracy with color-coded confidence band (Green/Yellow/Red)

Dynamic Field Customization

Add and Reprocess custom fields instantly using prompt-based inputs

Intelligence leverage LLM

Perform translation, Summarisation, Applied reasoning in case of needed

Intelligent Extraction

Use Cloud Document AI Intelligence Service (Both Azure/GCP) data from Structure and Unstructure Documents

Flexible Storage and API Based Processing and Retrieval

Store Results in MongoDB with unique processing_id.
Build a API with all above feature have flexibility to use different AI Services

Highlevel Flow

Upload Section

Upload Document (PDF/Image)

Upload

Upload Schema (JSON)

Upload

Extraction Results

| Field | Value |
|---------------|---------------------------------------|
| Name | Jane Smith |
| Date of Birth | 1980-03-25 |
| Address | 456 Oak Avenue, Springfield, IL 62701 |
| Amount | \$5.000.00 |
| Date | 04/01/2024 |
| ID Number | 12345 |
| Email | janesmith.email.com |

+ Add Custom Field

Validate

Next

Extraction Layer:
Leverage **Google Document AI / Azure Document Intelligence for Extraction** - Comes with Scoring and we can parse and share the confidence level

Translate

☒ Translation

Optional

French

▼

☐ Summary

Translation Results

| | |
|---------------------------------|-------------------|
| Invoice Number | INV-001 |
| Date de la facture | 15 janvier 2024 |
| Uplond chiema | Dupont SA |
| Ville | Paris |
| Montant total | 2 500,00 € |
| Date d'échéance | 30 janvier 2024 |
| Numéro de téléphone | +33-1-2345-6789 |
| Adresse électronique | bmartin#dupont.fr |
| + Ajouter un champ personnalisé | |

Summarize

Next

Intelligence Layer: Here you can use **LLM (Like GeminiPro)** to Translate, Summary, even we can apply some prompt for calculation and reasoning.

```
{
  "processing_id": "docproc_20250411_00123",
  "timestamp": "2025-04-11T17:45:00Z",
  "document_metadata": {
    "filename": "invoice_abc.pdf",
    "uploaded_by": "user@example.com",
    "uploaded_at": "2025-04-11T17:40:00Z",
    "schema_id": "schema_abc_001"
  },
  "extraction_results": {
    "fields": [
      {
        "name": "Invoice Number",
        "value": "INV-001",
        "confidence": 0.96,
        "status": "validated",
        "highlight": "green"
      },
      {
        "name": "Invoice Date",
        "value": "2024-02-12",
        "confidence": 0.93,
        "status": "validated",
        "highlight": "green"
      },
      {
        "name": "Amount",
        "value": "500.00",
        "confidence": 0.81,
        "status": "reviewed",
        "highlight": "amber"
      },
      {
        "name": "Email",
        "value": "johndoe@abc.com",
        "confidence": 0.38,
        "status": "manual_validation",
        "highlight": "red"
      }
    ],
    "custom_fields": []
  },
  "translation": {
    "language": "fr",
    "fields": [
      {
        "name": "Numéro de facture",
        "value": "INV-001"
      },
      {
        "name": "Date de la facture",
        "value": "12.02.2024"
      },
      {
        "name": "Client",
        "value": "ABC Corporation"
      },
      {
        "name": "Montant",
        "value": "500,00 €"
      },
      {
        "name": "Statut",
        "value": "En retard"
      },
      {
        "name": "Adresse e-mail",
        "value": "johndoe@abc.com"
      }
    ]
  },
  "summary": {
    "language": "fr",
    "text": "Le document est une facture émise par ABC Corporation, datée du 12.02.2024, d'un montant total de 500,00 €."
  },
  "settings": {
    "llm": "gemini-pro",
    "extractor": "google-document-ai",
    "storage_backend": "elasticsearch"
  },
  "status": "completed"
}
```

Completed: docproc_20250411_00123

Storage Layer: Each processing will contain a **processing_id** contains all steps and lineage of extraction, intelligence applied and store in **Document Database**.

Why this Hybrid Approach ?

| Criteria | End-to-End LLMs | Hybrid (Document AI for Extraction + LLM for NL Intelligence) | Enterprise Benefit for Hybrid Approach |
|--------------------------------|--|--|---|
| Accuracy & Structure Handling | Struggles with 2D layouts, inconsistent schema output | High accuracy on tables/forms with robust structure handling | Reliable for finance, legal, invoices, and compliance |
| Confidence Scoring | Lacks native scoring; difficult for human review workflows | Provides field-level scoring (Green/Amber/Red) | Enables human-in-the-loop (HITL) validation |
| OCR & Scanned Document Support | Requires external OCR | Built-in OCR with layout detection | Easier to deploy across scanned contracts and records |
| Translation Capability | Integrated but ungoverned | LLM-powered but schema-aligned, with human review options | Better multilingual control and governance |
| Integration Flexibility | Unified but less controllable | Modular – choose best tools (Google, Azure, Gemini), We can build enterprise required API for Downstream Consumption | Tailored to internal enterprise standards |
| Initial Setup | Simple to get started | Requires multi-service integration | One-time investment for long-term control |
| Scalability & Governance | Harder to control consistency at scale | Strong audit trails, schema enforcement, and governance | Scales safely with trust |
| Cost Efficiency | Lower initially but may create rework | Higher per document, but higher precision reduces manual rework | Cost justified by reduced errors and risk |