# Intelligent Document Processing System

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

## 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** 

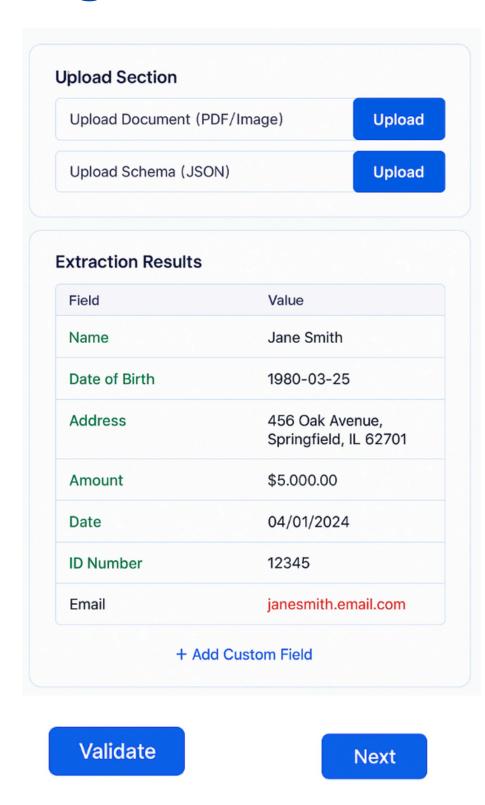
Use Cloud Document AI
Intelligence Service (Both
Azure/GCP) data from Structure
and Unstructure Documents

Flexible Storage and API Based Processing and Retreival

Store Results in MongoDB with unique processing\_id.

Build a API with all above feature have flexibility to use different AI Services

### Highlevel Flow

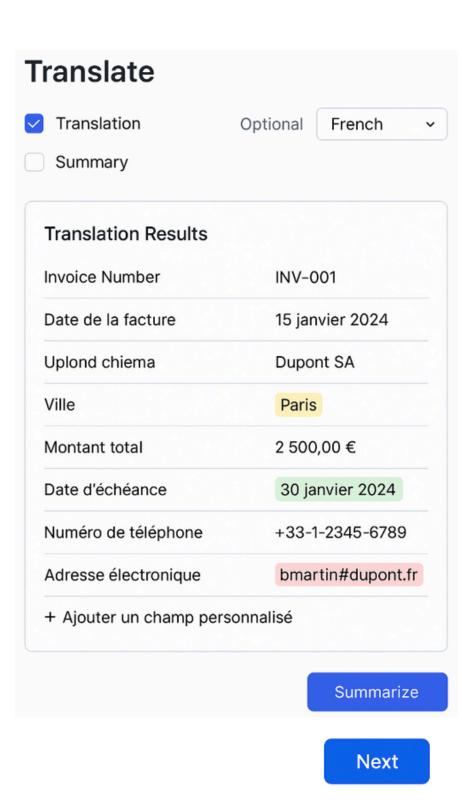


#### **Extraction Layer:**

Leverage Google Document AI / Azure

Document Intelligence for Extraction 
Comes with Scoring and we can parse

and share the confidence level



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

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

Completed: docproc\_20250411\_00123

Storage Layer: Each processing will contain a processing\_id contains all steps and lineage of extraction, intellegence 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