

1. AI for Data Searching / Extraction (PDF → JSON)

Free Tools

Unstructured.io (open-source) – extracts text, tables, and metadata from PDFs, docs, and images.

Camelot / Tabula – open-source PDF table extractors.

Tesseract OCR – free OCR for images/scanned docs.

OpenAI / HuggingFace LLMs – convert extracted text into JSON or structured fields.

2. AI for Data Cleaning

Free Tools

PandasAI (free) – uses LLMs on top of Pandas to clean and correct data.

Cleanlab (open-source) – identifies label errors, anomalies, and bad data.

Great Expectations (open-source) – validates and corrects inconsistent data.

OpenRefine (free) – cleans large messy datasets.

★ Use: Fix missing values, inconsistent units, noisy sensor data, incorrect labels.

3. AI for Data Structuring (Text → JSON/Schema)

Free Tools

Open-source LLMs (Llama 3, Mistral) – convert raw text into JSON structures.

Jina AI / FastEmbed (free) – embed and structure unstructured logs/text.

Pydantic (open-source) – validate and auto-structure data into Mongo-ready JSON.

LangChain (free) – pipeline to transform messy inputs → structured output.

★ Use: Convert PDF tables or logs into MongoDB-friendly JSON automatically.

4. AI for Data Injection / ETL into MongoDB

Free Tools

Airbyte (open-source) – automates file → database ingestion with transformations.

Dolt / Meltano (open-source) – ETL automation with Python transform steps.

FastAPI + open LLM – custom AI-powered transform → insert into MongoDB.

MongoDB Atlas Functions (free tier) – auto-triggered functions to clean and insert data.

★ Use: Upload files → auto-transform → auto-store cleaned JSON in MongoDB.