



UNSTRUCTURED

ETL for LLMs

What We Do

Whether you're training your own LLM, using a vector database, or pushing data through a ML pipeline, Unstructured effortlessly ingests and preprocesses natural language data.

Why It Matters

LLMs are trained on generic data, their knowledge is frozen in time, and without grounding to validated data, they're prone to hallucinations.

How We Work

Unstructured addresses these challenges with our comprehensive, enterprise-grade ETL solution that connects, transforms, and stages natural language data for LLMs.

Connectors to Any Location

Ingest Any File Type & Layout

- OneDrive
- MongoDB
- slack
- Google Drive
- salesforce
- elastic
- snowflake
- Azure
- databricks
- aws
- S
- box

- Audio
- EPUB
- JPEG/PNG
- MD
- MSFT Office
- PDF
- RST
- RTF
- TSV
- TXT
- XLS/CSV
- EML
- HTML
- OST
- PST
- SQL
- TAR
- Zip

- Auto
- Manual

Pipelines Optimize Cost/Accuracy

- Experimental
- High Res
- OCR Only
- Fast

Data Extracted/Transformed to Common Elements

JSON by Document Element

Generate Embeddings for Similarity Search

- Hugging Face
- Open AI
- Cohere

Deliver to Storage

- MongoDB Atlas
- Weaviate
- Pinecone
- chroma
- drant
- zilliz
- aws

Connect

Transform

Stage

Get Started Here

Contact



API Key



GitHub



CO

sales@unstructured.io