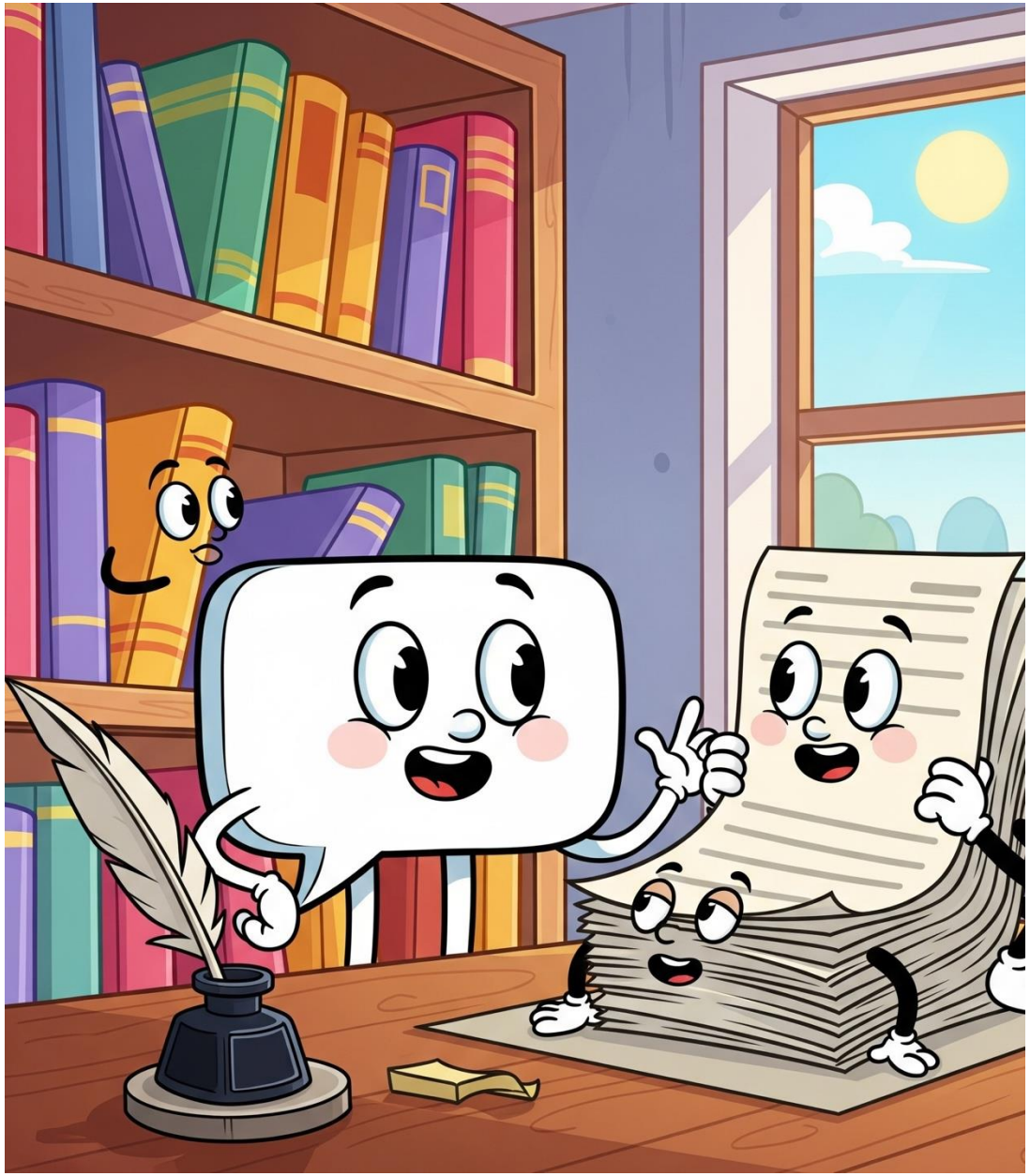
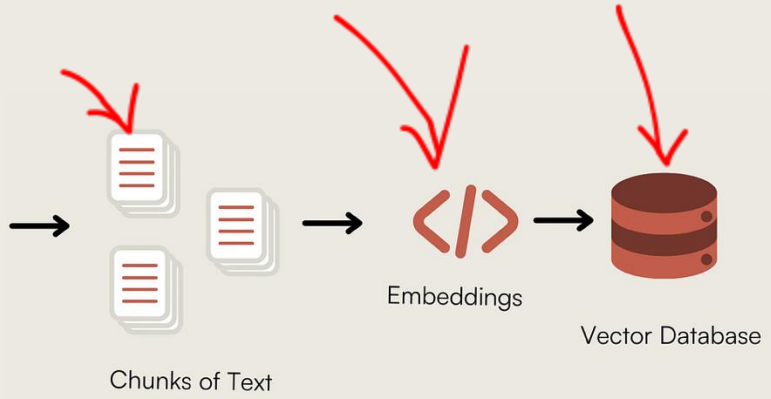
Chat with document? RAG?

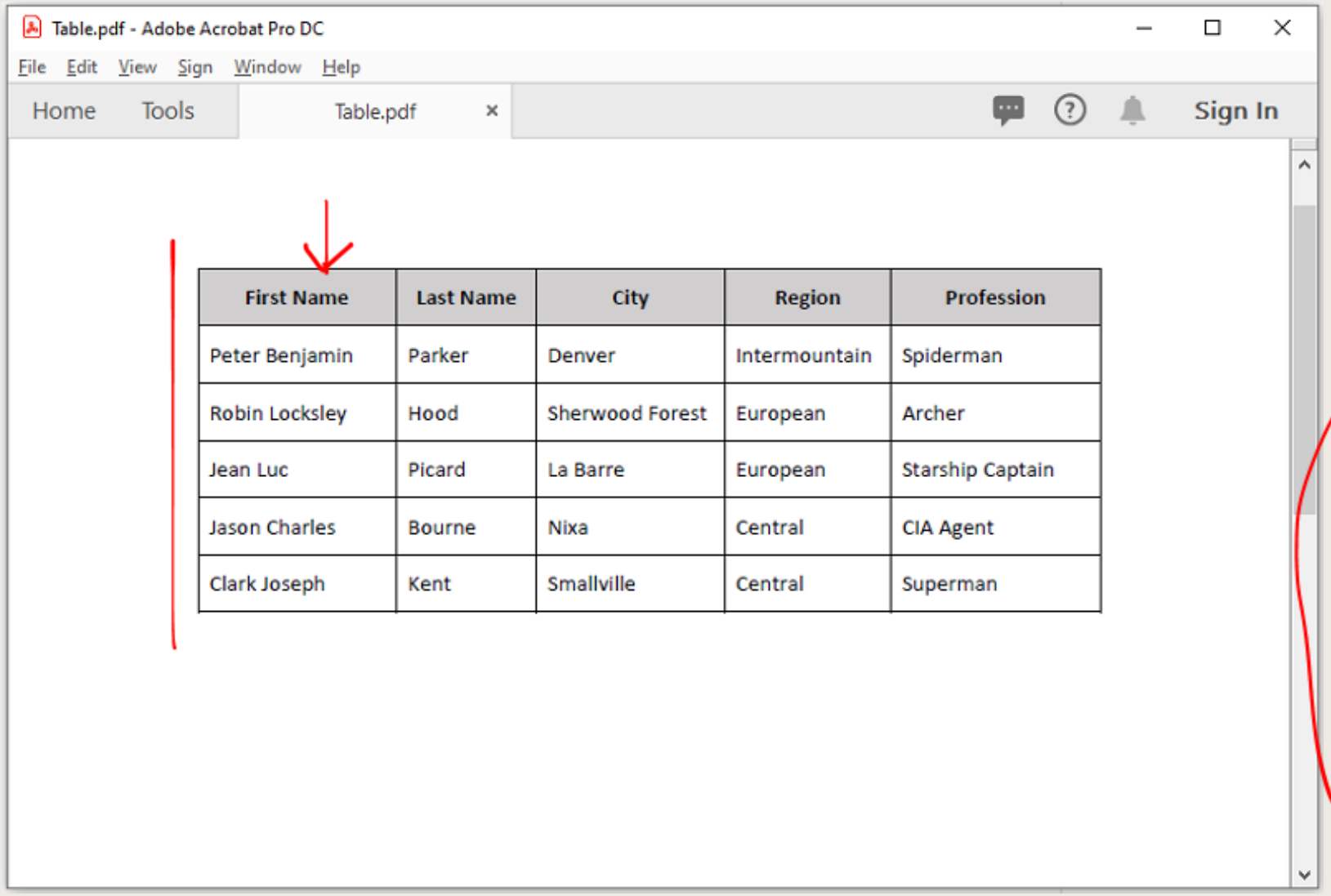


Write Everyday — Edited

Nov. 11

Call me Ishmael. Some years ago—never mind how long precisely—
having little or no money in my purse, and nothing particular to
interest me on shore, I thought I would sail about a little and
see the watery part of the world. It is a way I have of driving
off the spleen and regulating the circulation. Whenever I find
myself growing grim about the mouth; whenever it is a damp,
drizzly November in my soul; whenever I find myself involuntarily
pausing before coffin warehouses, and bringing up the rear of
every funeral I meet; and especially whenever my hypos get such an
upper hand of me, that it requires a strong moral principle to
prevent me from deliberately stepping into the street, and

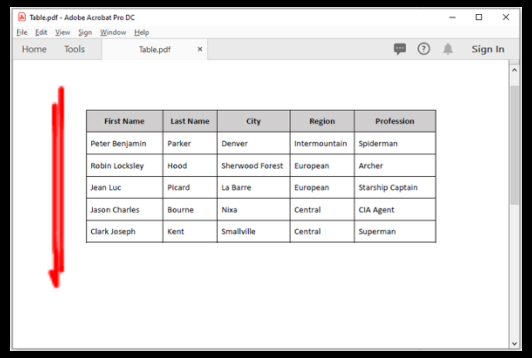




First Name	Last Name	City	Region	Profession
Peter Benjamin	Parker	Denver	Intermountain	Spiderman
Robin Locksley	Hood	Sherwood Forest	European	Archer
Jean Luc	Picard	La Barre	European	Starship Captain
Jason Charles	Bourne	Nixa	Central	CIA Agent
Clark Joseph	Kent	Smallville	Central	Superman

?

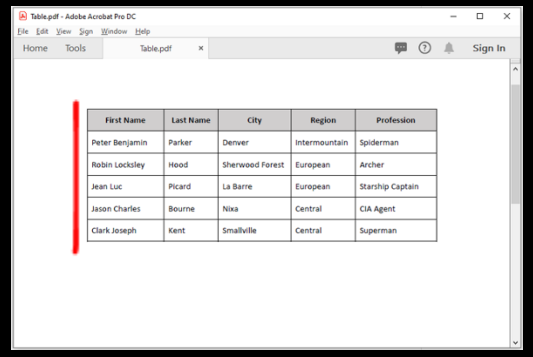
First Name Last Name City Region Profession Peter Benjamin Parker Denver Intermountain Spiderman Robin Locksley Hood Sherwood Forest European Archer Jean Luc Picard La Barre European Starship Captain Jason Charles Bourne Nixa Central CIA Agent Clark Joseph Kent Smallville Central Superman



markdown

Copy code

```
| First Name      | Last Name | City            | Region        | Profession       |
|-----------------|-----------|-----------------|---------------|------------------|
| Peter Benjamin  | Parker    | Denver          | Intermountain | Spiderman        |
| Robin Locksley  | Hood      | Sherwood Forest | European      | Archer           |
| Jean Luc        | Picard    | La Barre        | European      | Starship Captain |
| Jason Charles   | Bourne    | Nixa            | Central       | CIA Agent        |
| Clark Joseph    | Kent      | Smallville      | Central       | Superman         |
```

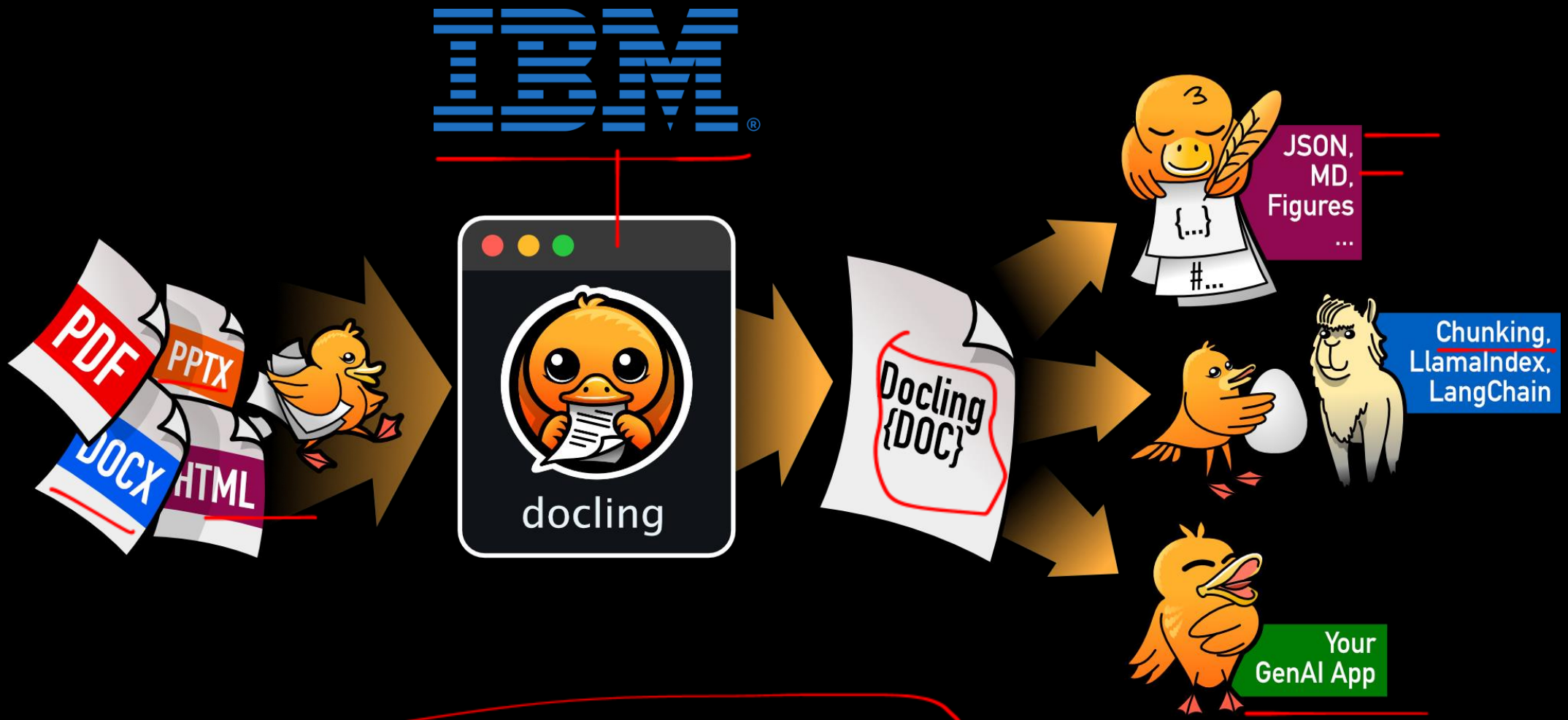


First Name	Last Name	City	Region	Profession
Peter Benjamin	Parker	Denver	Intermountain	Spiderman
Robin Locksley	Hood	Sherwood Forest	European	Archer
Jean Luc	Picard	La Barre	European	Starship Captain
Jason Charles	Bourne	Nixa	Central	CIA Agent
Clark Joseph	Kent	Smallville	Central	Superman

json

Copy code


```
[
  {
    "First Name": "Peter Benjamin",
    "Last Name": "Parker",
    "City": "Denver",
    "Region": "Intermountain",
    "Profession": "Spiderman"
  },
  {
    "First Name": "Robin Locksley",
    "Last Name": "Hood",
    "City": "Sherwood Forest",
    "Region": "European",
    "Profession": "Archer"
  },
  {
    "First Name": "Jean Luc",
    "Last Name": "Picard",
    "City": "La Barre",
    "Region": "European",
    "Profession": "Starship Captain"
  },
  {
    "First Name": "Jason Charles",
    "Last Name": "Bourne",
    "City": "Nixa",
    "Region": "Central",
    "Profession": "CIA Agent"
  },
  {
    "First Name": "Clark Joseph",
    "Last Name": "Kent",
    "City": "Smallville",
    "Region": "Central",
    "Profession": "Superman"
  }
]
```



https://github.com/docling-project/docling

Features

- 🗂️ Parsing of multiple document formats incl. PDF, DOCX, PPTX, XLSX, HTML, WAV, MP3, images (PNG, TIFF, JPEG, ...), and more
- 📑 Advanced PDF understanding incl. page layout, reading order, table structure, code, formulas, image classification, and more
- 🧬 Unified, expressive DoclingDocument representation format
- ↪️ Various export formats and options, including Markdown, HTML, DocTags and lossless JSON
- 🔒 Local execution capabilities for sensitive data and air-gapped environments
- 🤖 Plug-and-play integrations incl. LangChain, LlamaIndex, Crew AI & Haystack for agentic AI
- 🔍 Extensive OCR support for scanned PDFs and images
- 👓 Support of several Visual Language Models (SmolDocling)
- 🎙️ Support for Audio with Automatic Speech Recognition (ASR) models
- 💻 Simple and convenient CLI

1. Giới thiệu tổng quan

- **IBM Docling là gì?**
  → Một thư viện mã nguồn mở của IBM giúp trích xuất, chuẩn hóa và chuyển đổi tài liệu không cấu trúc (PDF, Word, PowerPoint, HTML, v.v.) thành dữ liệu có cấu trúc (JSON, Markdown, CSV, ...).
- Điểm mạnh: giữ nguyên định dạng bảng, hình ảnh, metadata và ngữ cảnh.

---

2. Vấn đề Docling giải quyết

- Doanh nghiệp đang có **khối lượng lớn tài liệu phi cấu trúc** (báo cáo, hợp đồng, tài liệu kỹ thuật...).
- Khó khăn khi cần tìm kiếm, phân tích hoặc đưa dữ liệu này vào hệ thống AI/BI.
- Docling giúp **tự động hóa** quá trình này, giảm thời gian xử lý thủ công, tăng độ chính xác.

3. Kiến trúc & Cách hoạt động

- Input: PDF, DOCX, PPTX, HTML, ...
- Xử lý: OCR (nếu cần), phân tích bố cục (layout analysis), trích xuất bảng, đoạn văn, metadata.
- Output: dữ liệu sạch, có cấu trúc (Markdown, JSON, CSV).

  👉 Có thể minh họa bằng sơ đồ **Input → Docling Engine → Structured Data**.

---

4. Tính năng nổi bật

- **Hỗ trợ đa dạng định dạng tài liệu**.
- **Preserve tables & layouts**: Giữ nguyên cấu trúc bảng và biểu đồ.
- **Metadata extraction**: Trích xuất tác giả, ngày, version, ...
- **Export linh hoạt**: Markdown, JSON, CSV → dễ tích hợp với LLM, BI tools.
- Mã nguồn mở (open source) trên GitHub, dễ mở rộng.

5. Ứng dụng thực tế

- Chuẩn hóa dữ liệu đầu vào cho **AI/LLM** (RAG, Chatbot tài liệu, Legal AI).
- Tự động hóa **báo cáo doanh nghiệp, compliance, hợp đồng**.
- **Ngân hàng & tài chính**: trích xuất dữ liệu từ báo cáo thường niên, hợp đồng tín dụng.
- **Chính phủ**: số hóa văn bản hành chính.

Hands-on

Macbook Air M2

Google Collab

GPU A6000
của **ThueGPU.vn**

Run & serve for N8N

ThueGPU.vn tặng

10 voucher 100K

Like & Subscribe và

Kéo xuống phần mô tả video & bình luận để nhập form!