OpenSearchCon EUROPE

Integrating Docling with OpenSearch for Advanced RAG and Agentic Applications

🎤 Cesar Berrospi Ramis, IBM

🗓️ April 16, 2026

🌍 Prague, Czechia

Introduction

OpenSearchCon EUROPE

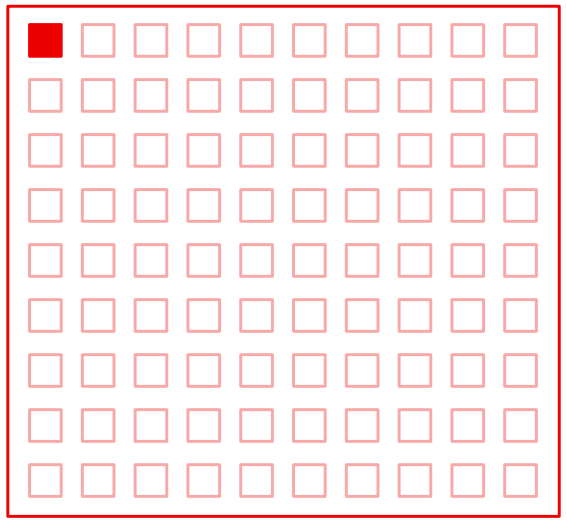
Cesar Berrospi Ramis

💼 Senior Research Scientist at IBM

🌐 https://ibm.biz/cesar-berrospi

The role of enterprise data

Knowledge/data cutoff for LLMs



Less than 1% of all enterprise data is represented in foundation models

Enterprise organizations need to

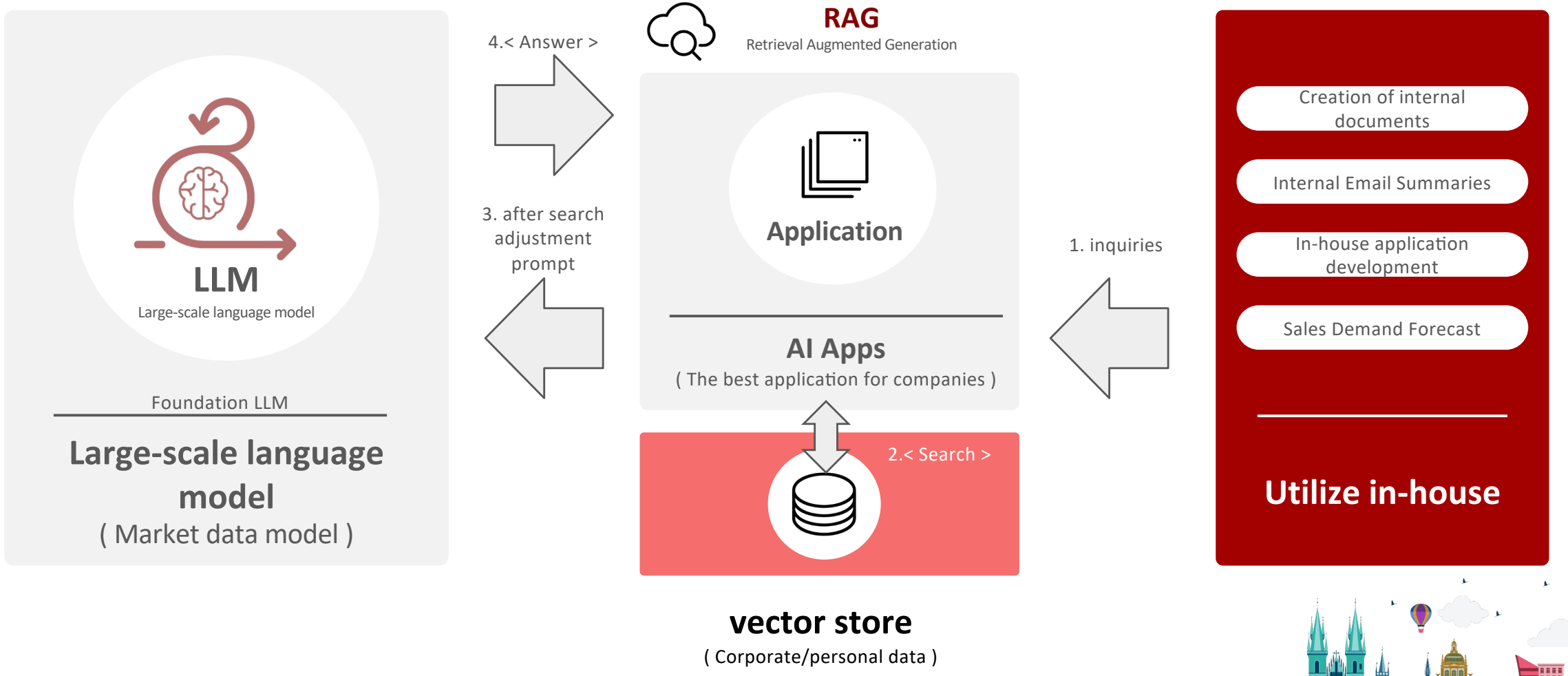
1. Start from a trusted base model
2. Create a new representation of their data(*)
3. Deploy, scale, and create value with their AI

* Much of the data is stuck in PDFs

Data extraction in RAG

OpenSearchCon EUROPE

RAG (Retrieval-Augmented Generation)




OpenSearch

AI Ready Search

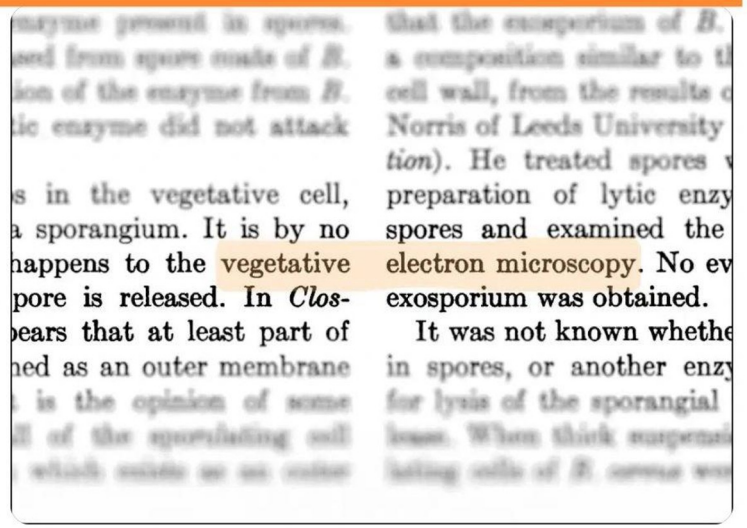
OpenSearchCon EUROPE

“Vegetative electron microscopy”

gurovdigital 15 h

lol, over 20 scientific papers now feature the nonsensical term ‘vegetative electron microscopy’.

all because an AI misinterpreted a 1959 article, merging ‘vegetative’ and ‘electron microscopy’ from separate columns.



...s in the vegetative cell, ...a sporangium. It is by no ...happens to the vegetative ...pore is released. In *Clos-* ...bears that at least part of ...ned as an outer membrane ... is the opinion of some ... of the sporulating cell ... which exists as an outer ... that the exosporium of *B.* ... a composition similar to t... cell wall, from the results ... Norris of Leeds University *tion*). He treated spores ... preparation of lytic enzy... spores and examined the ... electron microscopy. No ev... exosporium was obtained. It was not known whethe... in spores, or another enzy... for lysis of the sporangial ... loose. When thick suspens... lysing cells of *B. cereus* ...

692 12 43

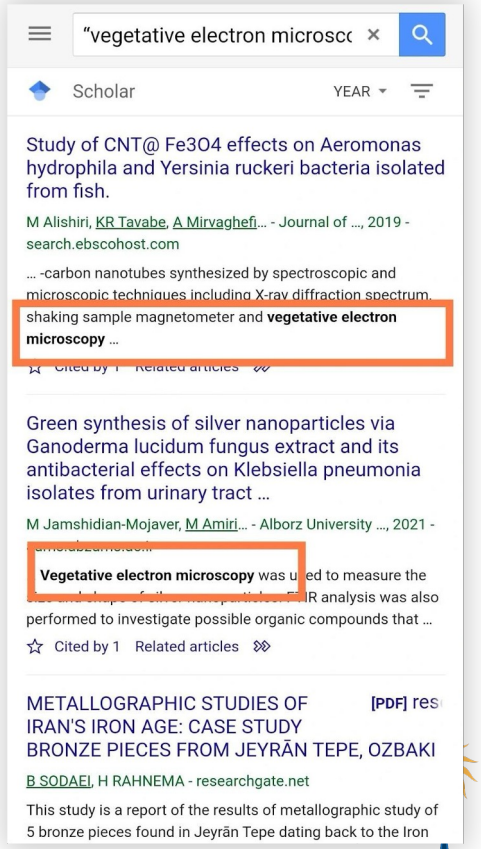
Date syrup (as one of the agricultural wastes) was used to produce bacterial cellulose using Gluconastobacter xylinus. Fourier transform infrared spectroscopy (FTIR), vegetative electron microscopy, and X-ray diffraction were used to determine the structure of bacterial cellulose, cellulose fibers, and crystallinity of the samples (Moosavi and Yousefi, 2011). After 14 days of incubation at 28 °C, the highest yield of cellulose

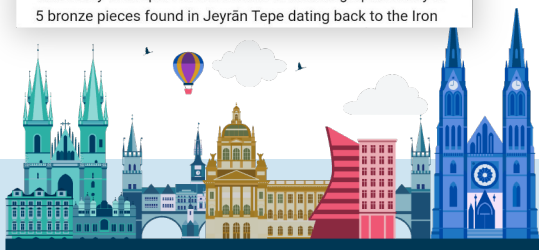
Silver and gold nanoparticles for antimicrobial purposes against multi-drug resistance bacteria [HTML] m

N Rabiee, S Ahmadi, O Akhavan, R Luque - Materials, 2022 - mdpi.com

... Dead bacteria have been observed by imaging and elemental analysis using transmission electron microscopes (TEM), **vegetative electron microscopy**, and EDX (X-Ray Probe Micro ...

☆ Cited by 112 Related articles »





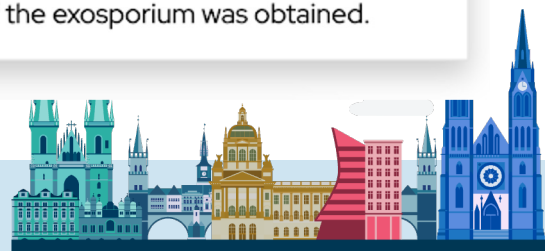
Well, actually… (sneak peek)

were incubated with an extract from spores dis-integrated at pH 7.0. Peptide was released which established that the coats contained sub-strate for the lytic enzyme present in spores. Peptide was also released from spore coats of *B. megaterium* by the action of the enzyme from *B. cereus* spores. The lytic enzyme did not attack intact resting spores.

The spore develops in the vegetative cell, which thus becomes a sporangium. It is by no means certain what happens to the vegetative cell wall when the spore is released. In *Clostridium* species it appears that at least part of this structure is retained as an outer membrane around the spore. It is the opinion of some workers that the wall of the sporulating cell forms the exosporium which exists as an outer

acteristic type. It was concluded that at least part of the sporangial wall was dissolved away to allow release of the spore. It appears likely that the exosporium of *B. cereus* does not have a composition similar to that of the vegetative cell wall, from the results obtained by Dr. J. R.

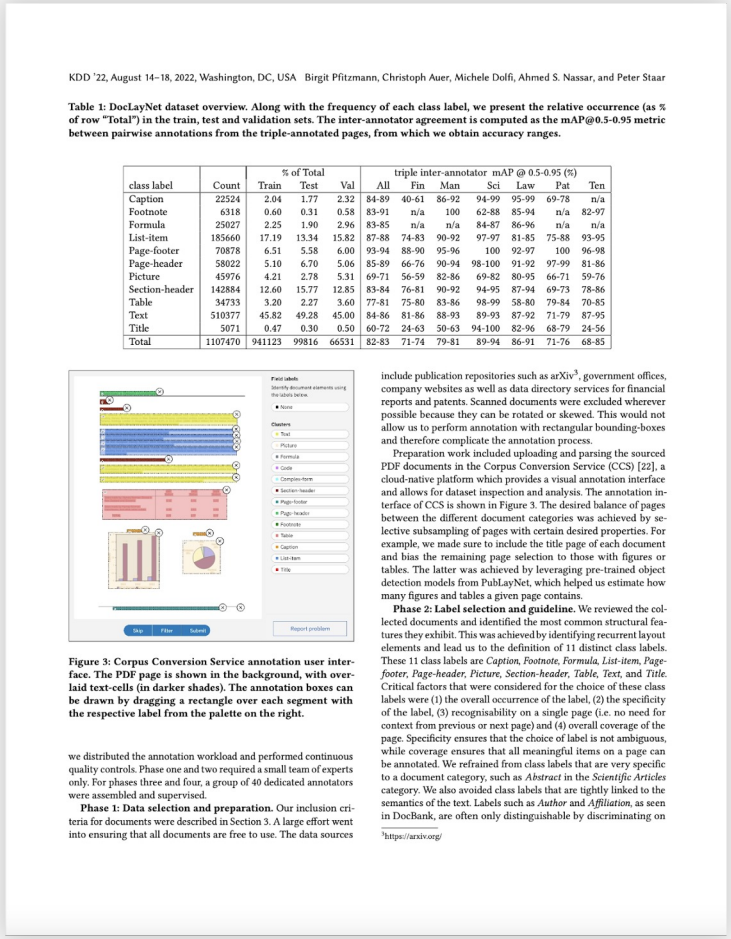
The spore develops in the vegetative cell, which thus becomes a sporangium. It is by no means certain what happens to the vegetative cell wall when the spore is released. In Clostridium species it appears that at least part of this structure is retained as an outer membrane around the spore. It is the opinion of some workers that the wall of the sporulating cell forms the exosporium which exists as an outer coat around spores of several Bacilus species. Spores of several varieties of B. cereus had exosporia whereas these structures appeared to be absent from spores of B. megaterium and B. subtilis. It seems, however, that in Bacillus species at least, the greater part of the vegetative cell wall is dissolved away before the developed spore is released. If this is true, then soluble components containing the characteristic constituents should appear in the medium during spore release. Culture filtrates from B. cereus organisms at various stages of growth and sporulation were hydrolyzed and the hydrolyzates analyzed for amino sugars and diaminopimelic acid (28). Results showed that a large increase in the concentration of these substances in the culture filtrate occurred during spore release (table 2); they were found to be present in a nondialyzable peptide of the characteristic type. It was concluded that at least part of the sporangial wall was dissolved away to allow release of the spore. It appears likely that the exosporium of B. cereus does not have a composition similar to that of the vegetative cell wall, from the results obtained by Dr. J. R. Norris of Leeds University (personal communication). He treated spores with a highly active preparation of lytic enzyme from B. cereus spores and examined the effect by means of electron microscopy. No evidence of lysis of the exosporium was obtained.



AI Ready Search

OpenSearchCon EUROPE

Structured content in PDFs



KDD '22, August 14–18, 2022, Washington, DC, USA Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter Staar

Table 1: DocLayNet dataset overview. Along with the frequency of each class label, we present the relative occurrence (as % of row "Total") in the train, test and validation sets. The inter-annotator agreement is computed as the mAP@0.5-0.95 metric between pairwise annotations from the triple-annotated pages, from which we obtain accuracy ranges.

		% of Total			triple inter-annotator mAP @ 0.5-0.95 (%)						
class label	Count	Train	Test	Val	All	Fin	Man	Sci	Law	Pat	Ten
Caption	22524	2.04	1.77	2.32	84-89	40-61	86-92	94-99	95-99	69-78	n/a
Footnote	6318	0.60	0.31	0.58	83-91	n/a	100	62-88	85-94	n/a	82-97
Formula	25027	2.25	1.90	2.96	83-85	n/a	n/a	84-87	86-96	n/a	n/a
List-item	185660	17.19	13.34	15.82	87-88	74-83	90-92	97-97	81-85	75-88	93-95
Page-footer	70878	6.51	5.58	6.00	93-94	88-90	95-96	100	92-97	100	96-98
Page-header	58022	5.10	6.70	5.06	85-89	66-76	90-94	98-100	91-92	97-99	81-86
Picture	45976	4.21	2.78	5.31	69-71	56-59	82-86	69-82	80-95	66-71	59-76
Section-header	142884	12.60	15.77	12.85	83-84	76-81	90-92	94-95	87-94	69-73	78-86
Table	34733	3.20	2.27	3.60	77-81	75-80	83-86	98-99	58-80	79-84	70-85
Text	510377	45.82	49.28	45.00	84-86	81-86	88-93	89-93	87-92	71-79	87-95
Title	5071	0.47	0.30	0.50	60-72	24-63	50-63	94-100	82-96	68-79	24-56
Total	1107470	941123	99816	66531	82-83	71-74	79-81	89-94	86-91	71-76	68-85

Figure 3: Corpus Conversion Service annotation user interface. The PDF page is shown in the background, with overlaid text-cells (in darker shades). The annotation boxes can be drawn by dragging a rectangle over each segment with the respective label from the palette on the right.

we distributed the annotation workload and performed continuous quality controls. Phase one and two required a small team of experts only. For phases three and four, a group of 40 dedicated annotators were assembled and supervised.

Phase 1: Data selection and preparation. Our inclusion criteria for documents were described in Section 3. A large effort went into ensuring that all documents are free to use. The data sources include publication repositories such as arXiv[3], government offices, company websites as well as data directory services for financial reports and patents. Scanned documents were excluded wherever possible because they can be rotated or skewed. This would not allow us to perform annotation with rectangular bounding-boxes and therefore complicate the annotation process.

Preparation work included uploading and parsing the sourced PDF documents in the Corpus Conversion Service (CCS) [22], a cloud-native platform which provides a visual annotation interface and allows for dataset inspection and analysis. The annotation interface of CCS is shown in Figure 3. The desired balance of pages between the different document categories was achieved by selective subsampling of pages with certain desired properties. For example, we made sure to include the title page of each document and bias the remaining page selection to those with figures or tables. The latter was achieved by leveraging pre-trained object detection models from PubLayNet, which helped us estimate how many figures and tables a given page contains.

Phase 2: Label selection and guideline. We reviewed the collected documents and identified the most common structural features they exhibit. This was achieved by identifying recurrent layout elements and lead us to the definition of 11 distinct class labels. These 11 class labels are *Caption, Footnote, Formula, List-item, Page-footer, Page-header, Picture, Section-header, Table, Text,* and *Title.* Critical factors that were considered for the choice of these class labels were (1) the overall occurrence of the label, (2) the specificity of the label, (3) recognisability on a single page (i.e. no need for context from previous or next page) and (4) overall coverage of the page. Specificity ensures that the choice of label is not ambiguous, while coverage ensures that all meaningful items on a page can be annotated. We refrained from class labels that are very specific to a document category, such as *Abstract* in the *Scientific Articles* category. We also avoided class labels that are tightly linked to the semantics of the text. Labels such as *Author* and *Affiliation,* as seen in DocBank, are often only distinguishable by discriminating on

[3] https://arxiv.org/

KDD '22, August 14–18, 2022, Washington, DC, USA Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Nassar, and Peter Staar

Table 1: DocLayNet dataset overview. Along with the frequency of each class label, we present the re occurrence (as %
of row "Total") in the train, test and validation sets. The inter-annotator agreement is computed as th 0.95 metric
between pairwise annotations from the triple-annotated pages, from which we obtain accuracy rang

% of Total

triple inter-annotator mAP @ 0.5-0.95 (%)

[...]
Count
22524
6318
25027
185660
70878
58022
45976
142884
34733
510377
5071
1107470

[...]

include publication repositories such as arXiv3, government o"ces, company websites as well as data directory services for #nancial reports and patents. Scanned documents were excluded wherever possible because they can be rotated or skewed. This would not allow us to perform annotation with rectangular bounding-boxes and therefore complicate the annotation process.

[...]

! undesired page headers

! Tables not understood

! Image content missing

! Line wraps not understood

! Multi-column often breaks order

✓ Very fast and cheap

✗ Incomplete

✗ Loss of structure

✗ Noisy

➞ Unfit for most use cases

What is Docling?

- Multiple parsing / export formats
- Advanced PDF understanding
- Unified `DoclingDocument` representation format
- Many plug-and-play ecosystem integrations
- Local execution for sensitive data

LF AI & DATA

Docling in action

OpenSearchCon EUROPE

2206.01062.pdf
Page 1 of 9

Search

DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis

Birgit Pfitzmann
IBM Research
Rueschlikon, Switzerland
bpf@zurich.ibm.com

Christoph Auer
IBM Research
Rueschlikon, Switzerland
cau@zurich.ibm.com

Michele Dolfi
IBM Research
Rueschlikon, Switzerland
dol@zurich.ibm.com

Ahmed S. Nassar
IBM Research
Rueschlikon, Switzerland
ahn@zurich.ibm.com

Peter Staar
IBM Research
Rueschlikon, Switzerland
taa@zurich.ibm.com

ABSTRACT

Accurate document layout analysis is a key requirement for high-quality PDF document conversion. With the recent availability of public, large ground-truth datasets such as PubLayNet and DocBank, deep-learning models have proven to be very effective at layout detection and segmentation. While these datasets are of adequate size to train such models, they severely lack in layout variability since they are sourced from scientific article repositories such as PubMed and arXiv only. Consequently, the accuracy of the layout segmentation drops significantly when these models are applied on more challenging and diverse layouts. In this paper, we present *DocLayNet*, a new, publicly available, document-layout annotation dataset in COCO format. It contains 80863 manually annotated pages from diverse data sources to represent a wide variability in layouts. For each PDF page, the layout annotations provide labelled bounding-boxes with a choice of 11 distinct classes. DocLayNet also provides a subset of double- and triple-annotated pages to determine the inter-annotator agreement. In multiple experiments, we provide baseline accuracy scores (in mAP) for a set of popular object detection models. We also demonstrate that these models fall approximately 10% behind the inter-annotator agreement. Furthermore, we provide evidence that DocLayNet is of sufficient size. Lastly, we compare models trained on PubLayNet, DocBank and DocLayNet, showing that layout predictions of the DocLayNet-trained models are more robust and thus the preferred choice for general-purpose document-layout analysis.

CCS CONCEPTS

• **Information systems** → **Document structure**; • **Applied computing** → **Document analysis**; • **Computing methodologies** → **Machine learning**; **Computer vision**; *Object detection*;

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3539043

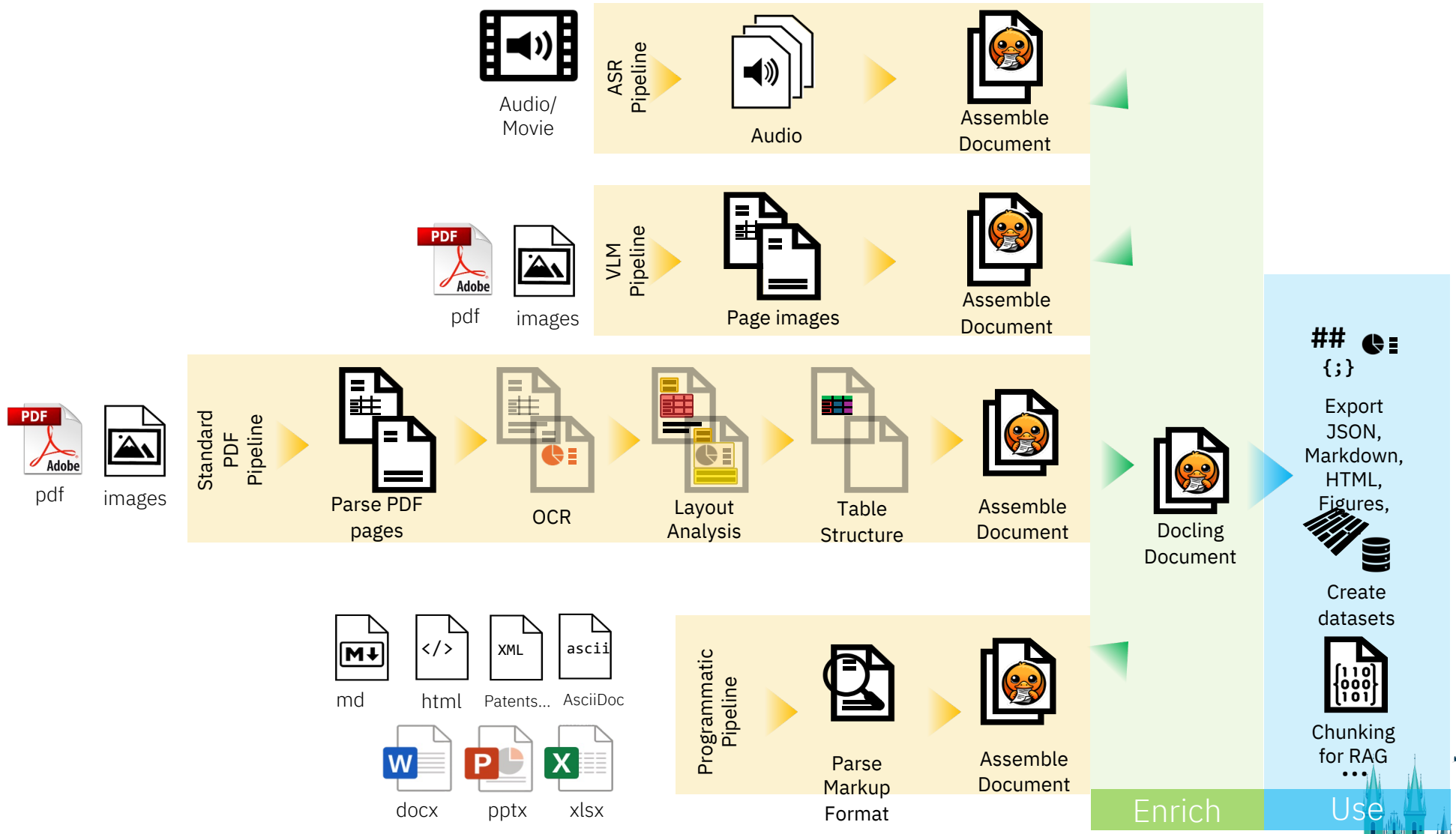
Figure 1: Four examples of complex page layouts across different document categories

KEYWORDS

PDF document conversion, layout segmentation, object-detection, data set, Machine Learning

ACM Reference Format:
Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter Staar. 2022. DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539043

Docling Architecture



How to use Docling

Docling CLI

```
> pip install docling
```

Installation options: https://docling-project.github.io/docling/getting_started/installation/

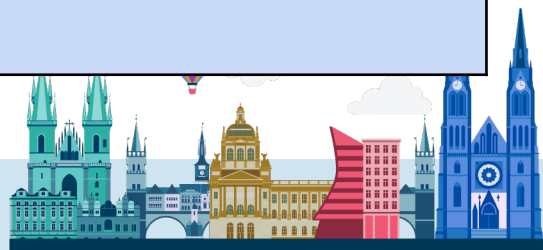
```
> docling path/to/horrible.pdf
```

CLI reference: https://docling-project.github.io/docling/reference/cli/

How to use Docling

Docling in Python

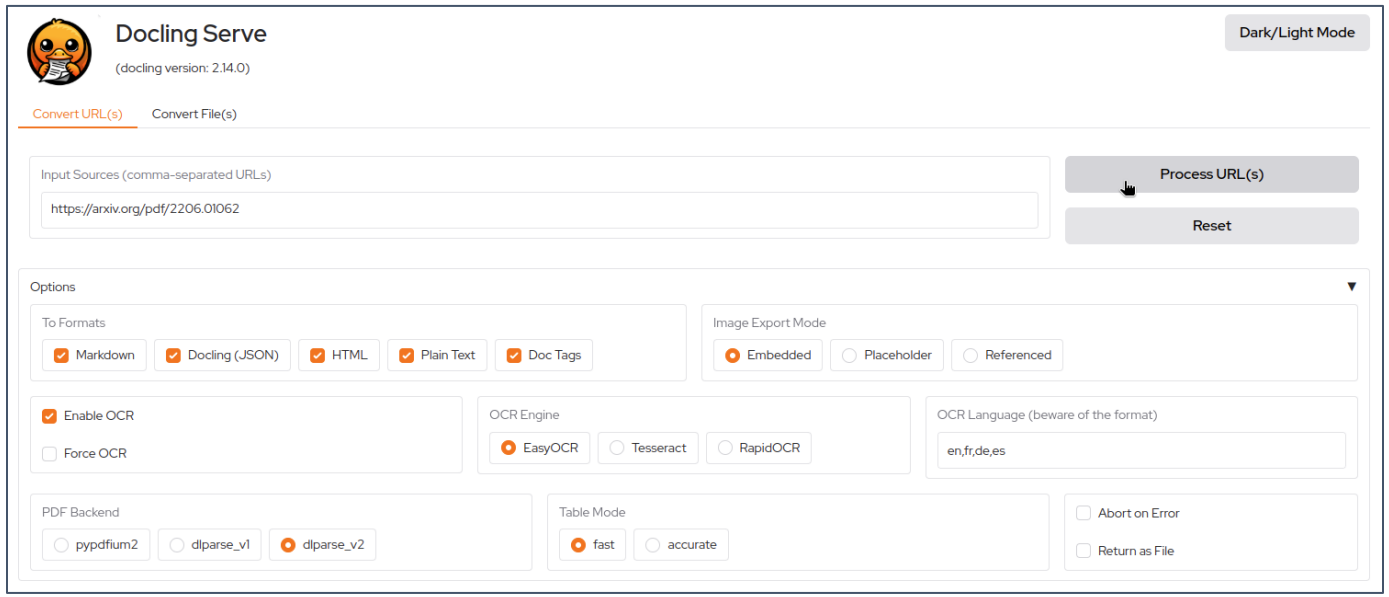
```
from docling.document_converter import DocumentConverter

source = "https://arxiv.org/pdf/2408.09869"
converter = DocumentConverter()
doc = converter.convert(source).document
print(doc.export_to_markdown())
```



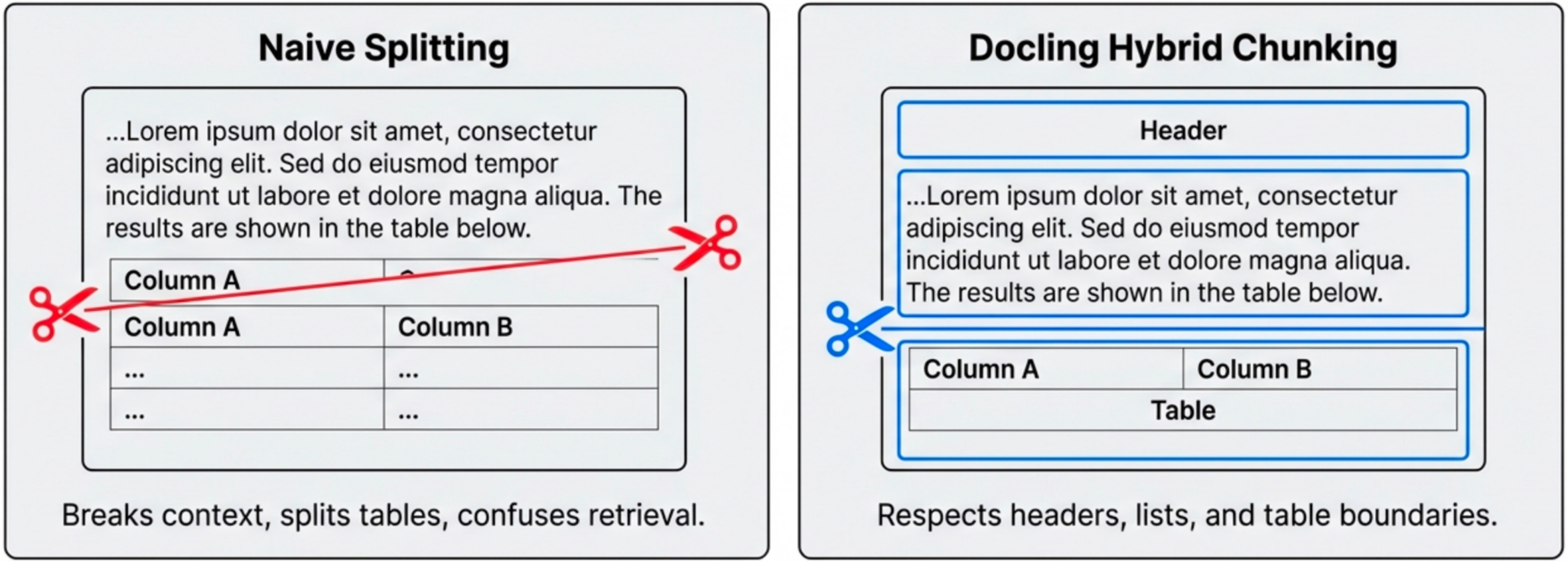
How to use Docling

Docling as an API service (and UI)

```
> pip install docling-serve
> docling-serve run --enable-ui
```



Docling Hybrid Chunker



AI Ready Search

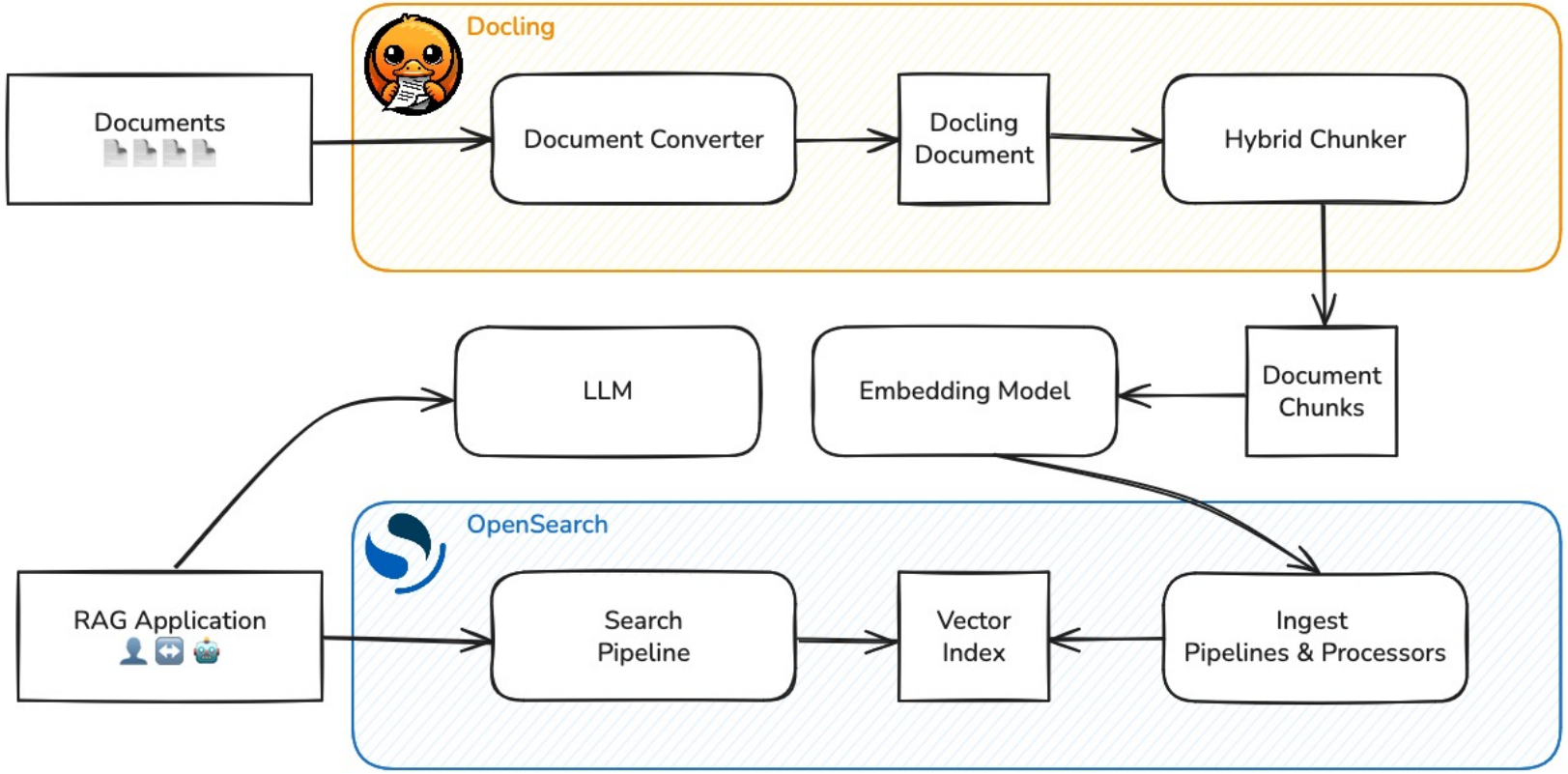
What do we need for AI ready search?

Text extracted (images described, table structure retained)

Turned into chunks

Indexed in OpenSearch

A simple RAG application



https://github.com/docling-project/docling/blob/main/docs/examples/rag_opensearch.ipynb

Tighter integration with Docling Java API

A custom ingestion plugin leverages Docling Java API to process several document types into a rich, unified format (**DoclingDocument**) that can be used in OpenSearch ingestion pipelines (expected release: 2Q 2026)

OpenSearch Cluster

Ingest Node

Ingest-Docling Plugin

IngestDoclingPlugin
- Registers DoclingProcessor.Factory

DoclingProcessor
- Receives IngestDocument
- Extracts document from field
- Calls DoclingImpl.parse()
- Populates target field with results

DoclingImpl
- Security wrapper with restricted permissions
- Creates DoclingServeApi client
- Sends document bytes to Docling Serve
- Receives DoclingDocument response
- Extracts markdown and metadata
- Returns DoclingResult

Docling Java Client Library

DoclingServeApi
- HTTP client for Docling Serve
- Handles request/response serialization
- Built with Java HttpClient + Jackson

Data Models
- ConvertDocumentRequest (with BytesSource)
- ConvertDocumentResponse (with DoclingDocument)
- DoclingDocument (Java representation)

Docling Serve Instance (External Service)

Docling Serve API
- Receives document bytes via REST API
- Routes to appropriate document converter

Docling Core Processing Engine
- PDF (with layout analysis, table detection,...) parser
- DOCX,PPTX,XLSX,HTML,...parsers

DoclingDocument (Unified Format)
- Document structure and content
- Layout information
- Tables, images, formulas

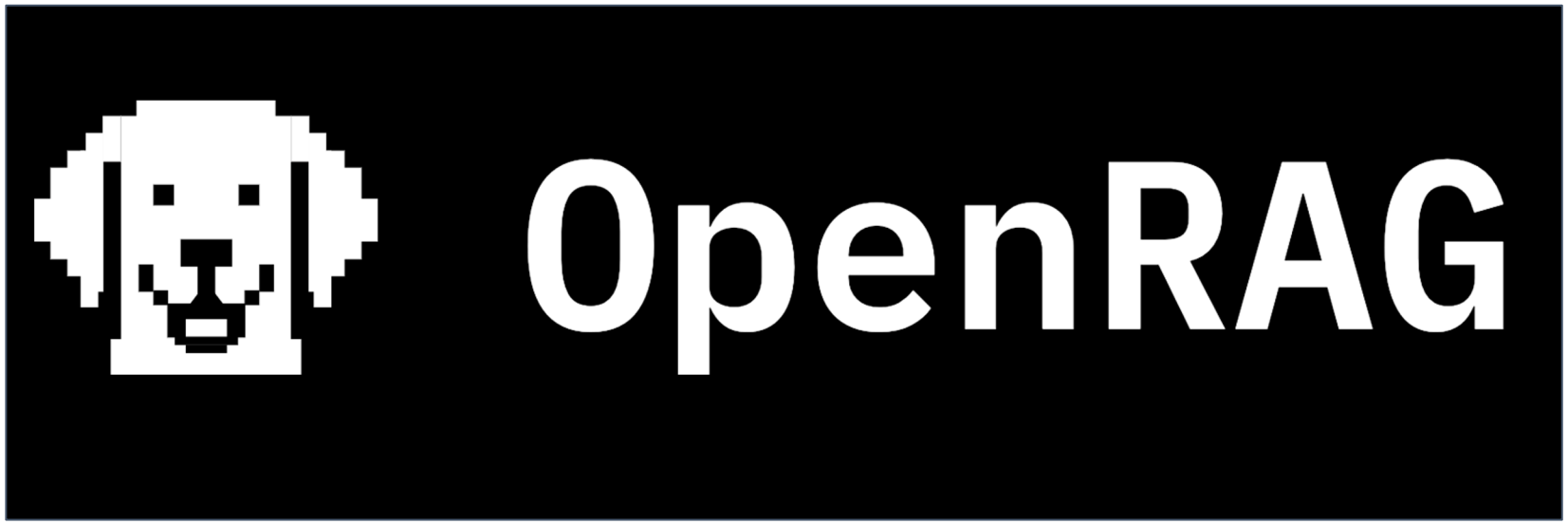
Export to Markdown
- Converts DoclingDocument to Markdown format
- Preserves structure (headings, lists, tables)
- Returns formatted text

Agentic search with Docling MCP

Docling provides an MCP server with tools for document conversion, manipulation, or content creation

Integration examples with several frameworks and AI desktop clients

With OpenSearch native support for external MCP tools, run Docling using only OpenSearch API

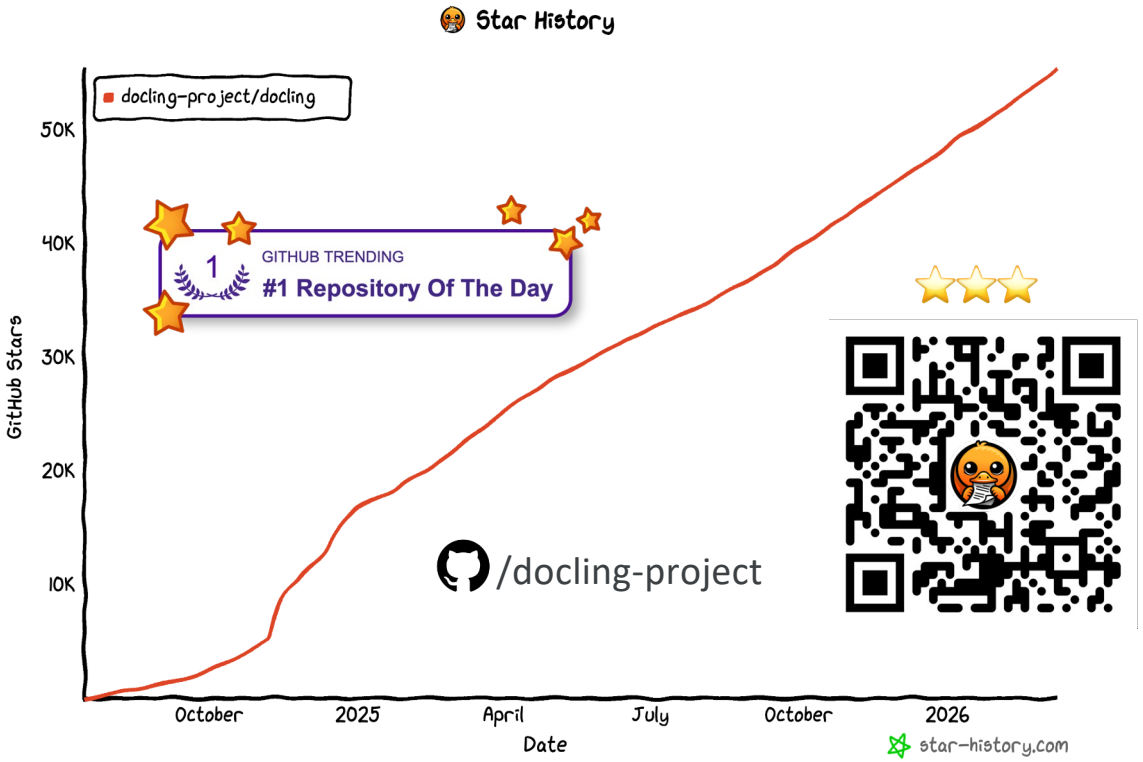


LANGFLOW OPENSEARCH DOCLING

https://github.com/langflow-ai/openrag

Community adoption

- Community adoption
- ⭐ 55k+ GitHub stars
- 🐍 4.7M+ downloads last month from PyPI



Recent highlights

OpenSearchCon EUROPE

Docling Java project

NVIDIA x Docling collab

NVIDIA DEVELOPER Home Blog Forums Docs Downloads Training

Technical Blog Search blog

An Agentic AI toolkit for local AI

The use cases for private, local agents are endless. But building reliable, repeatable, and high-quality private agents remains a challenge. LLM quality deteriorates when you distill and quantize the model to fit within a limited VRAM budget on PC. The need for accuracy increases as agentic workflows require reliable and repeatable answers when interfacing with other tools or actions.

To address this, developers typically use two tools to increase accuracy: fine-tuning and retrieval-augmented-generation (RAG). NVIDIA released updates to accelerate tools across this workflow for building agentic AI.

Nemotron 3 Nano is a 32B parameter MoE model optimized for agentic AI and fine-tuning. With 3.6B active parameters and a 1M context window, it tops several benchmarks across coding, instruction-following, long-context reasoning, and STEM tasks. The model is optimized for RTX PCs and DGX Spark via Ollama and llama.cpp, and can be fine-tuned using Unsloth.

This model stands out for being the most open, with weights, recipes, and datasets widely available. Open models and datasets make customizing the model easier for developers. They prevent redundant fine-tuning and eliminate data leakage for objective benchmarking for robust and efficient workflows. Get started with LoRA-based fine-tuning for it.

For RAG, NVIDIA partnered with Docling—a package to ingest, analyze, and process documents into a machine-understandable language for RAG pipelines. Docling is optimized for RTX PCs and DGX Spark and delivers 4x performance compared to CPUs.

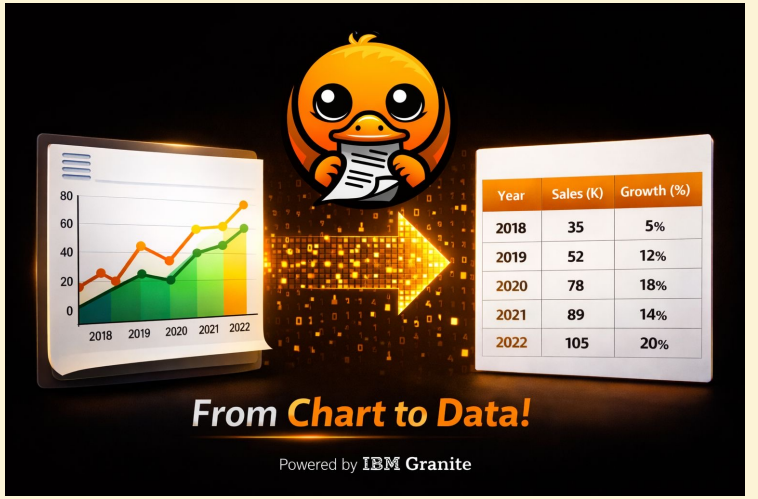
There are two ways of using Docling:

1. Traditional OCR pipeline: This is a pipeline of libraries and models that is accelerated via PyTorch-CUDA on RTX.
2. VLM-based pipeline: An advanced pipeline for complex multi-modality documents, available for use via vLLM within WSL and Linux environments.

Docling is developed at IBM and contributed to the Linux Foundation. Start now on RTX with this easy-to-use guide.

LaTeX parsing

Chart understanding



Coming up for Docling

- Model runtimes and presets
- Metrics packages
- Memory hardening and profiling
- Enhanced large scale processing
- XBRL input support
- docling-slim package
- Docling-as-a-Service

https://lfaidata.foundation/projects/docling/


OpenSearchCon
EUROPE
docling.ai