DocSpiral: A Platform for Integrated Assistive Document Annotation through Human-in-the-Spiral

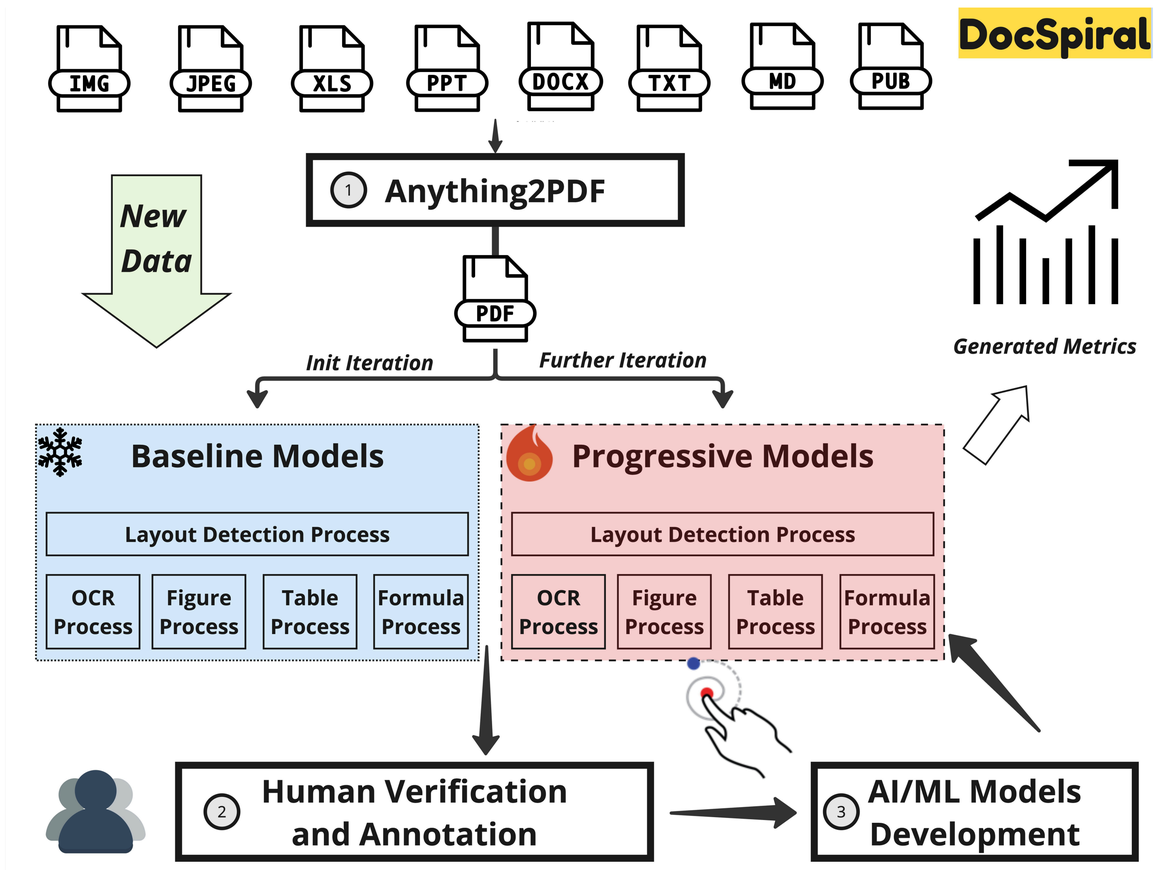
Qiang Sun*, Sirui Li, Tingting Bi, Du Huynh, Mark Reynolds, Yuanyi Luo, Wei Liu*
* Corresponding Author

Motivation

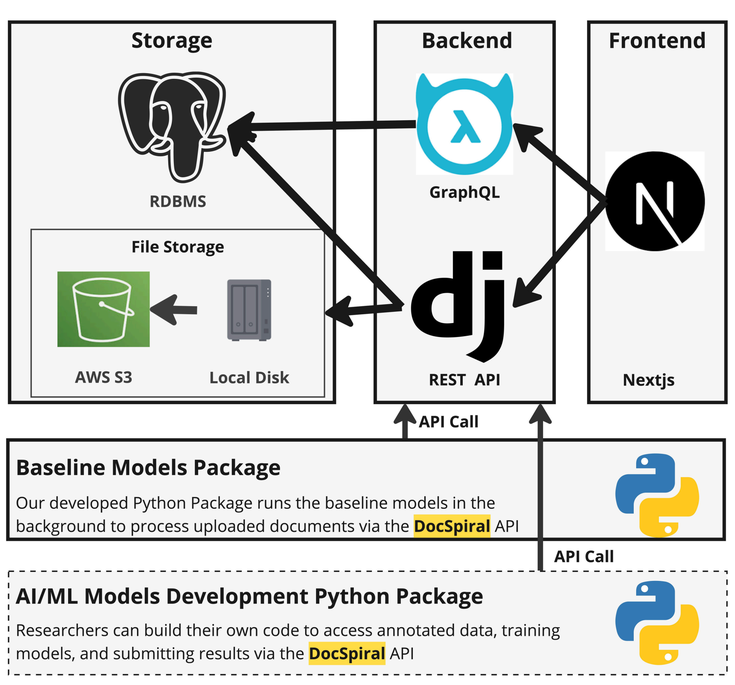
To support the development of domain-specific technical document AI/ML models, we need a **Comprehensive Annotation Platform.**

- **Domain-specific challenges**: Techical documents such as geological exploration reports, hospital discharge letters, invoices need specialized models due to domain specific layouts and terminology, etc
- **Data accessibility crisis**: 80-90% of valuable knowledge are trapped in scanned PDFs and images
- **Tool fragmentation:** Existing annotation tools only support partial workflows, not end-to-end pipelines

Concept



System Design

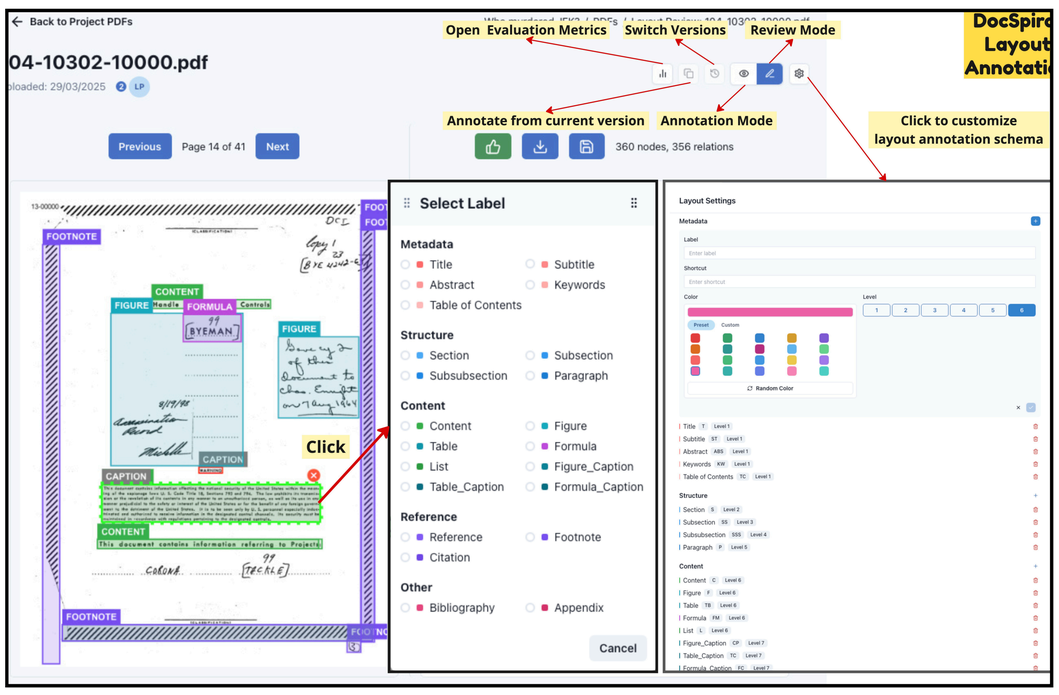


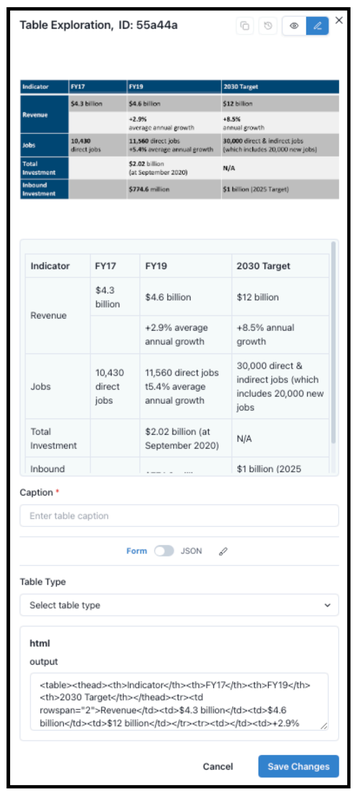
Tools Comparison

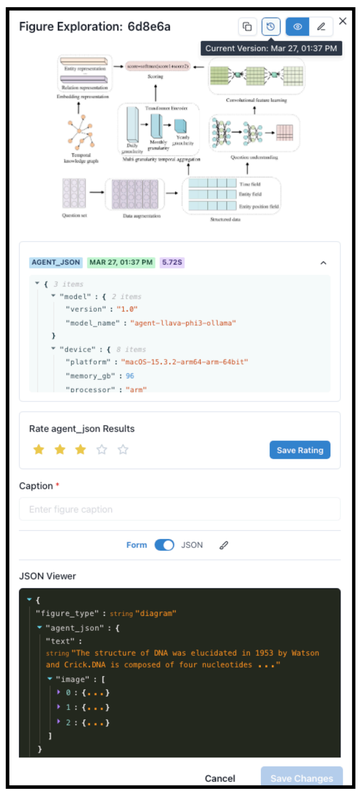
Table 1: Comparison of Document Annotation Tools. Due to the emergence of LLM and RAG technologies still being a recent development, tools supporting figure, formula, and table understanding capabilities remain scarce. (*Ann. ⇒ Annotation, Conv. ⇒ Conversion: transforming data from image to another while preserving complete factual content without interpretation, Und. ⇒ Understanding: generating descriptive text based on a given image, involving interpretation, meaning inference, pattern recognition, and subjective judgment about data implications.*)

Tool	Reference	Open Access	Layout Ann.	OCR Ann.	Figure		Formula		Table	
					Conv.	Und.	Conv.	Und.	Conv.	Und.
ABBYY FineReader	(ABBYY, 1993)	No	×	✓	×	×	×	×	✓	✓
Transkribus	(READ-COOP SCE, 2013)	No	×	✓	×	×	×	×	✓	×
Coco Annotator	(Brooks, 2019)	Yes	✓	×	×	×	×	×	×	×
PDFAnno	(Shindo et al., 2018)	Yes	×	✓	×	×	×	×	×	×
Label Studio	(Tkachenko et al., 2020)	Partially	✓	✓	×	×	×	×	×	×
PPOCRLabelv2	(PFCCLab, 2020)	Yes	✓	✓	×	×	×	×	✓	✓
PAWLS	(Neumann et al., 2021)	Yes	✓	×	×	×	×	×	×	×
Tagtog	(TagTog team, 2023)	No	×	✓	×	×	×	×	×	×
Prodigy	(Explosion AI, 2023)	No	✓	×	×	×	×	×	×	×
Callico	(Kermorvant et al., 2024)	No	×	✓	✓	✓	×	×	×	×
DocSpiral	Ours	Yes	✓	✓	✓	✓	✓	✓	✓	✓

Feature Highlights







Empirical Study

https://app.ai4wa.com

https://nlp-tlp.org

Annotation Efficiency Improvements

41%	Overall Time Reduction
75%	Reduction for Low-Quality PDFs
90	Document Pages Tested

Model Performance Evolution

Faster-RCNN Layout Detection Training Results

Metric	Initial	1st	2nd	3rd
mAP (%)	5.3 →	12.0 →	21.0 →	33.0
Progress	—	+6.7%	+9.0%	+12.0%

Each iteration represents one complete cycle of the annotation-training spiral
Progressive improvement demonstrates the effectiveness of human-in-the-loop training

SCAN ME

ACL 2025 VIENNA
The 63rd Annual Meeting of the Association for Computational Linguistics
July 27–August 1st, 2025, Vienna, Austria