



Uni-Parser Technical Report

DP Technology *

December 16, 2025

<https://uni-parser.github.io>
 <https://huggingface.co/UniParser>

Abstract

This technical report introduces *Uni-Parser*, an industrial-grade document parsing engine tailored for scientific literature and patents, delivering high throughput, robust accuracy, and cost efficiency. Unlike pipeline-based document parsing methods, Uni-Parser employs a modular, loosely coupled multi-expert architecture that preserves fine-grained cross-modal alignments across text, equations, tables, figures, and chemical structures, while remaining easily extensible to emerging modalities. The system incorporates adaptive GPU load balancing, distributed inference, dynamic module orchestration, and configurable modes that support either holistic or modality-specific parsing. Optimized for large-scale cloud deployment, Uni-Parser achieves a processing rate of up to 20 PDF pages per second on 8 × NVIDIA RTX 4090D GPUs, enabling cost-efficient inference across billions of pages. This level of scalability facilitates a broad spectrum of downstream applications, ranging from literature retrieval and summarization to the extraction of chemical structures, reaction schemes, and bioactivity data, as well as the curation of large-scale corpora for training next-generation large language models and AI4Science models.

1 Introduction

The rapid advancement of large language models (LLMs) has significantly expanded the scope of document-centric applications, ranging from intelligent assistants and domain-specific agents to automated knowledge base construction. Central to these developments is the ability to reliably parse and structure information from PDF documents, which remain the dominant medium for disseminating scientific knowledge. High-quality structured data extracted from scientific literature is particularly critical for downstream LLM applications, enabling accurate reasoning, retrieval-augmented generation, and decision-making across a wide spectrum of scientific and industrial tasks.

Among scientific domains, chemical and biomedical literature holds exceptional importance and immense commercial value. Parsing such documents enables the creation of comprehensive molecular and reaction databases, bioactivity repositories, material structure archives, and experimental characterization datasets. These resources not only accelerate drug discovery and materials design but also serve as the foundation for emerging AI4Science research. However, despite decades of progress in OCR, table recognition, and layout analysis, the majority of the hundreds of millions of scientific and patent PDFs remain underexploited. This is largely because the PDF format, while optimized for human readability, poses extraordinary challenges for computational processing: extraction is prohibitively costly.

Current approaches, both pipeline-based [1–4] and VLM-based [5–11], face three major challenges. First, they are computationally inefficient, making it prohibitively costly to parse tens of millions of

* Full authorship contribution statements appear at the end of the document.

documents at scale. Second, their performance on non-textual modalities such as formulas, tables, charts, and chemical structures remains limited, with low accuracy and poor robustness. Third, the complex and heterogeneous layouts of scientific and patent documents often lead to unreliable segmentation and structural analysis, further degrading downstream usability. Pipeline methods offer higher throughput but poor generalization, while VLM methods generalize better but suffer from low efficiency and limited extensibility. Collectively, these limitations hinder the construction of large-scale, high-quality knowledge bases required for both academic research and industrial innovation.

To address these challenges, we present *Uni-Parser*, an industrial-grade, multi-modal PDF parsing engine purpose-built for scientific literature and patents. Uni-Parser follows the pipeline-based methods, and combines high throughput with state-of-the-art accuracy through a modular, loosely coupled architecture composed of specialized expert models for different modalities. The system introduces a set of key innovations:

- **High-efficiency large-scale inference:** A distributed microservice design with dynamic GPU load balancing enables real-time parsing throughput, supporting fast and cost-effective inference over billions of document pages.
- **Accurate multi-modal parsing:** A suite of domain-specialized, lightweight expert models achieves state-of-the-art accuracy across text, equations, tables, figures, and chemical structures.
- **Robust layout recognition for scientific and patent documents:** A newly designed layout analysis and reading order algorithm tailored to complex publishing formats greatly improves reliability in handling dense, irregular, and domain-specific page structures.

Together, these contributions establish Uni-Parser as a scalable and extensible foundation for structured document understanding. By transforming unstructured PDFs into clean, machine-actionable representations, Uni-Parser not only supports immediate applications such as literature retrieval, summarization, and knowledge extraction, but also enables the construction of domain-specific repositories in chemistry, materials science, and biomedicine—paving the way for data-driven AI4Science innovation.

2 Algorithm Framework

2.1 Overall Framework

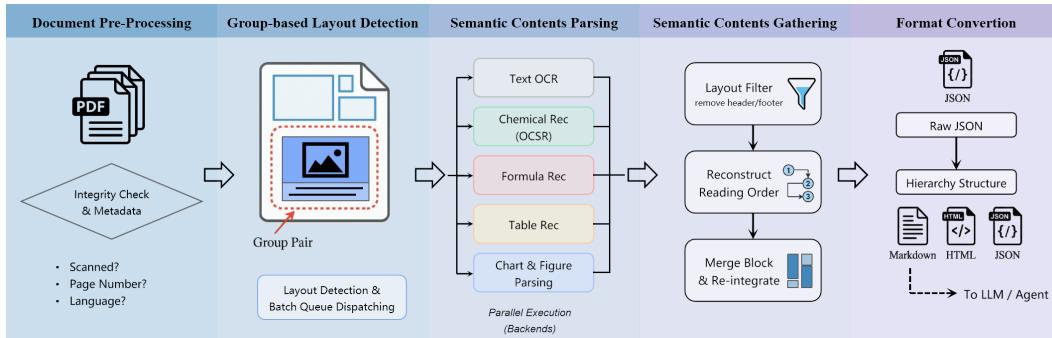


Figure 1: Sketch of the Uni-Parser pipeline. Uni-Parser converts unstructured PDFs into clean, hierarchical, multimodal outputs (text, formulas, tables, figures, and chemical structures). These enriched representations are designed to be readily consumed by large language models, enabling more accurate understanding, reasoning, and document-level operations.

Uni-Parser adopts a modular design with a strong emphasis on extensibility. As shown in the figure, it first conducts validation and pre-processing on PDF file inputs. It then converts these unstructured documents into machine-readable formats (e.g., JSON) or LLM-compatible representations (e.g., Markdown and HTML) through a sequence of models and processing stages. These stages are organized into five principal components:

Document Pre-Processing: During pre-processing, the system ingests PDFs from user uploads or URLs, verifies file integrity and encryption, and extracts metadata such as page count, dimensions, and text accessibility. Documents with corrupted or garbled text are marked as non-extractable. If embedded text layers exist, they are directly extracted; otherwise, a lightweight OCR method provides partial text for language identification. Uni-Parser supports over 80 languages for OCR mode. This stage typically takes around one hundred milliseconds and supports parallel processing of multiple PDFs.

Group-based Layout Detection: The layout detection model locates each semantic block in the PDF pages and identifies its semantic category. Unlike conventional approaches [2, 3], our group-based layout detection model recognizes naturally paired semantic components—such as image–caption, table–title, and molecule–identifier—preserving visual–semantic associations that are critical for structured information extraction and accurate reading-order construction. Finally, different semantic blocks are rendered at dynamic resolutions and forwarded to designated models for image-to-text processing. We employ a greedy batch-stacking strategy for rendering page images, performing layout recognition, and dispatching semantic groups to downstream microservices. As soon as a batch accumulates in the queue, it is immediately processed by subsequent modules, allowing the latency of this stage to be almost entirely masked by other processing steps.

Semantic Contents Parsing: Semantic contents are parsed by routing each block to the appropriate specialized model. General text blocks are processed with OCR, tables with table recognition models, mathematical formulas with formula recognition models, molecular structures with optical chemical structure recognition (OCSR), chemical reactions with dedicated reaction extraction models, and charts with chart parsing models. In total, over ten sub-models and specialized procedures handle diverse content types, including text, tables, molecular structures, chemical reactions, formulas, and charts. All models operate in parallel to maximize efficiency. For image blocks, the system can retain either visual descriptions or raw image data, depending on downstream application requirements. This stage is the most time-consuming stage.

Semantic Contents Gathering: Building on the parsed results from the previous stage, this phase filters out non-essential elements such as headers and footers while preserving key combinations like figure–caption or table–caption pairs and horizontal separators. Content blocks within each page are reordered to reflect the logical structure of the document. Multimodal elements embedded in text lines or table cells—such as inline equations, chemical structures, or charts—are reintegrated into their corresponding text or table. Cross-page and multi-column content, including tables, paragraphs, and reaction schemes, is merged to maintain coherence. The system also incorporates the original PDF’s section hierarchy to guide the final organization. The output of this stage provides a fully structured, semantically enriched representation of the document, suitable for downstream analysis and reading-order reconstruction.

Output Formatting and Semantic Chunking: To support diverse downstream tasks, the fully parsed document can be exported in task-specific formats, or as a complete plain text, interleaved image–text, Markdown, or HTML. Thanks to group-based layout detection and the merging of semantic content across columns and pages, reconstructed paragraphs or semantic groups are output as properly chunked data, which improves semantic coherence. This approach streamlines chunking and facilitates more efficient downstream processing, such as retrieval-augmented generation (RAG). Furthermore, PDF section headings are integrated into the document structure when available. These metadata elements enrich the final representation with hierarchical navigation cues, improving both user-facing applications and LLM-driven analysis.

2.2 Group-based Layout Analysis

Layout detection is a foundational prerequisite for all subsequent document analysis tasks and represents a pivotal component within document parsing systems. Its efficacy ultimately establishes the performance ceiling for the entire system. Acknowledging the profound diversity and complexity of scientific document layouts, along with the critical importance of grouping relationships among heterogeneous semantic elements, we introduce a layout understanding paradigm that diverges significantly from conventional approaches [1, 2, 12, 13]: a group-based tree-structured layout representation.

Operationally, we conceptualize the layout of a document page as a hierarchical organization, enabling the aggregation of semantically related elements into coherent and logical groups. For instance, figures are paired with their captions, tables with their titles, equations with their reference numbers, and molecular structures with their identifiers, among other relational pairings. Figure 2 shows an example of the layout tree structure. Although document layouts inherently involve multiple levels of hierarchy, in our annotation we restrict the structure to only two layers: the bottom layer and the top layer. The bottom layer covers all fundamental semantic components, serving as the parent nodes of the layout structure tree, whereas the top layer comprises content nested within the bottom layer or other top level semantic elements, functioning as child nodes. The corresponding semantic categories are summarized in Table 1. Importantly, this design still allows downstream post-processing to recover the nested semantic content across multiple hierarchical levels. Based on these concepts, we constructed a group-based layout detection model named *Uni-Parser-LD*.

Table 1: Layout type used in Uni-Parser-LD

Layout Type	Layout Layer	Role	Semantic Parsing Category
Document Title	Bottom Layer	Main Text & Structure Role	OCR
Section Title		Main Text & Structure Role	OCR
Paragraph		Main Text	OCR
References		Supplementary Text	OCR
Table of Contents		Supplementary Text	OCR
Key-value Item		Main Text	OCR
Code Block		Figure / Multi-modal Text	OCR / Code Parsing
Header	Bottom Layer	Functional Role	–
Footer		Functional Role	–
Footnote		Supplementary Text	OCR
Sidebar		Functional Role	–
Page Number		Functional Role	OCR
Watermark		Functional Role	–
Divider Line		Functional Role	–
Formula ¹	Bottom Layer	Multi-modal Text	Formula Recognition
Table ²		Multi-modal Text	Table Structure Recognition
Image ³		Figure / Multi-modal Text	Image Caption
Formula (Inline)	Top Layer	Multi-modal Text	Formula Recognition
Molecule ⁴		Multi-modal Text	OCSR
Chemical Reaction ⁵		Figure / Multi-modal Text	Reaction Parsing
Chart ⁵		Figure / Multi-modal Text	Chart to Table
Figure ⁵		Figure / Multi-modal Text	Image Caption

¹ Grouped with formula ID.

² Grouped with table caption and table footnote.

³ Grouped with image caption.

⁴ Grouped with molecule identifier and Markush description.

⁵ Grouped with figure legend and figure caption.

Considering the wide variety of authoring and rendering styles across layouts and modalities, we employ a large-scale dataset of PDF pages for training. We built an in-house layout detection dataset containing 500k pages. Of these, 220k pages are carefully human annotated with group-based layout labels from a diverse corpus. The database primarily consists of scientific journal data and patent data from various patent offices. It also includes preprints, books, and other types of data across different fields, spanning a total of 85 languages. For more details, refer to Section 4. The remaining pages are synthetic data used for pretraining. Our experiments show that when high-quality, large-scale real data is available, pretraining with existing synthetic datasets, such as DocSynth300K [12], can be counterproductive. This is due to their low-fidelity rendering, stylistic deviations from authentic documents, and limited diversity imposed by manually defined generation rules, which fail to capture the complexity of real-world layouts. To address this, we generate synthetic samples from real layouts using controlled modifications—including element merging, spatial perturbations, and semantic content substitutions—allowing the model to effectively leverage synthetic data for pretraining.

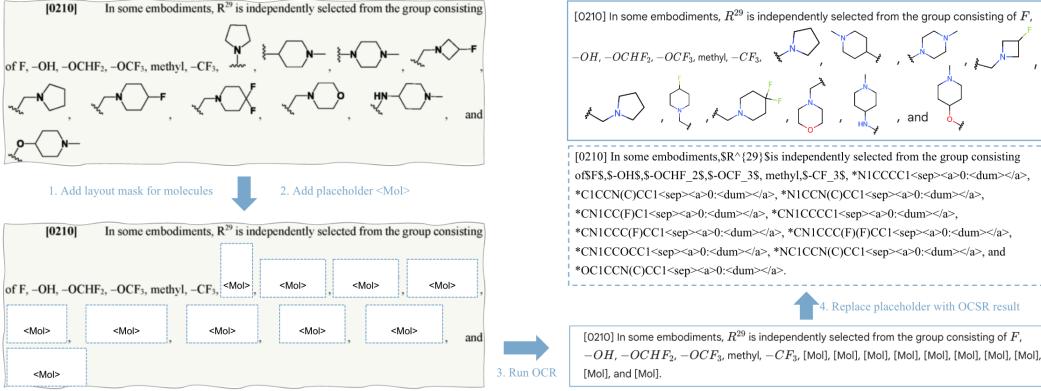


Figure 3: An example of the OCR model inference workflow in Uni-Parser. When a top-layer layout element overlaps a bottom-layer layout block, the system substitutes it with a placeholder before performing OCR. The placeholder is then resolved during post-processing, enabling fast and accurate multimodal parsing.

2.4 Table Structure Recognition

Tables represent compact yet structurally complex layouts, as each cell may contain heterogeneous semantic blocks, including plain text, mathematical formulas, molecular structures, reaction schemes, images, or even nested sub-tables. The presence of such multi-level nesting makes it challenging for a single end-to-end model to perform generalized OCR on tables. To address this, we adopt a modular strategy that decouples table structure recognition from multimodal content parsing. This approach not only improves interpretability but also enhances robustness and overall performance.

Before structure recognition, Uni-Parser recovers the orientation of tables, since rotated layouts frequently occur in patents and supplementary materials of scientific articles, particularly when tables are extended across pages. Instead of employing a dedicated orientation model, we leverage layout predictions and auxiliary metadata to perform lightweight orientation recovery, achieving accuracy comparable to that of specialized models while significantly improving efficiency.

We adopt an efficient table structure recognition model, SLANet [20], trained on a corpus of one million tables that integrates a cleaned version of PubTabNet [21], SynthTabNet [22], and our in-house synthetic dataset of line-based tables enriched with multimodal content. The synthetic dataset covers both bordered and borderless formats and incorporates diverse elements, including molecule structures, charts, images, formulas, and text. This comprehensive training strategy allows SLANet to achieve strong generalization and superior performance on complex real-world table layouts.

	MOLSTRUCTURE	COMPOUND NAME	IC50
95		3-(2-CHLORO-BENZOYL)-METHYL-AMINO-5-PHENYL-THIOPHENE-2-CARBOXYLIC ACID	++
96		3-(2-METHYL-BENZOYLAMINO)-5-PHENYL-THIOPHENE-2-CARBOXYLIC ACID	++
97		3-(METHYL-(2-METHYL-BENZOYL)-AMINO)-5-PHENYL-THIOPHENE-2-CARBOXYLIC ACID	+

Figure 4: An example of table structure recognition results produced by Uni-Parser. By decoupling table structure recognition from table content recognition, the system achieves improved robustness, supports multimodal nesting within tables.

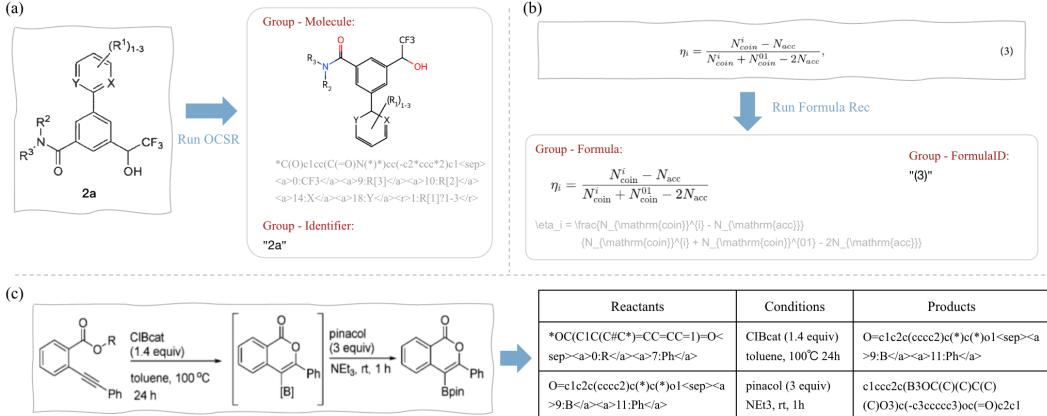


Figure 5: Examples of multi-modal recognition results produced by Uni-Parser. (a) Molecular structures are correctly associated with their corresponding identifiers. (b) Mathematical formulas are accurately linked to their formula IDs. (c) An organic chemical reaction is parsed into a structured reactant–condition–product triplet.

2.5 Formula Recognition

The system integrates a mathematical formula recognition module fine-tuned from PP-Formula [2]. The module converts mathematical expressions and chemical equations into LaTeX sequences in an end-to-end manner. It handles both standalone formulas and inline formulas or equations within paragraph text. For standalone formulas, the layout stage groups the expression with its reference number to form a structured representation.

2.6 Chemical Expression Recognition

Chemical structures are a fundamental modality in scientific literature and patents, particularly in chemistry, pharmaceuticals, biology, and materials. Uni-Parser adopts the end-to-end OCSR architecture MolParser [23] for molecular recognition, which translates molecular images into an Extended SMILES (E-SMILES) representation. We further introduce *MolParser 1.5*, which extends MolParser with an expanded pretraining corpus. In addition to the synthetic MolParser-7M dataset, we add a real-world dataset of 10 million images pseudo-labeled via cross-validation using multiple MolParser models and a fine-tuned MolScribe [24]. With a larger proportion of in-the-wild chiral molecules and Markush structures, MolParser 1.5 yields a more balanced and comprehensive pretraining dataset, and after fine-tuning on the MolParser-SFT dataset, it produces a more robust end-to-end OCSR model.

For chemical reaction image recognition, we adopt a pipeline approach. First, we detect the texts and molecular structures within the reaction in layout stage. Then, we identify the associations among blocks to construct a reactant–condition–product graph. Finally, we aggregate the results to obtain the parsed reaction equation.

2.7 Chart and Scientific Figure Understanding

Uni-Parser incorporates an optional module for chart and scientific figure understanding. When activated, the system converts charts into their underlying data tables. For figures that are not amenable to accurate tabular representation—such as spectra or complex scientific diagrams—the system generates detailed image captions that convey essential information, including key numerical values.

For chart understanding, we fine-tuned a Qwen-2.5-VL-3B [25] model on a dataset consisting of 500k generated samples, 300k open-source samples, and 170k real-world charts. This training significantly enhances the model’s ability to convert charts into tables, with a particular focus on multi-subplot charts, which frequently appear in scientific literature.

To support accurate scientific figure captioning, we curated a large-scale, high-quality dataset of 4 million samples. The core of this dataset consists of 3 million scientific images meticulously collected from high-impact publications. To enhance caption quality, we leveraged multiple multi-modal large language models (MLLMs), including but not limited to GPT-5 [26] and Gemini 2.5 [27], to rewrite the original captions by integrating visual content with raw captions and relevant contextual information from the associated papers. The dataset is further augmented with additional sources, including ChemPile [28], molecular descriptions from PubChem, Electron Microscopy image captions [29], and other experimental characterization datasets. Leveraging this comprehensive dataset, we fine-tuned a Qwen-2.5-VL-3B model, which named as *SciParser*, enabling it to generate precise, context-aware captions for a diverse range of scientific figures.

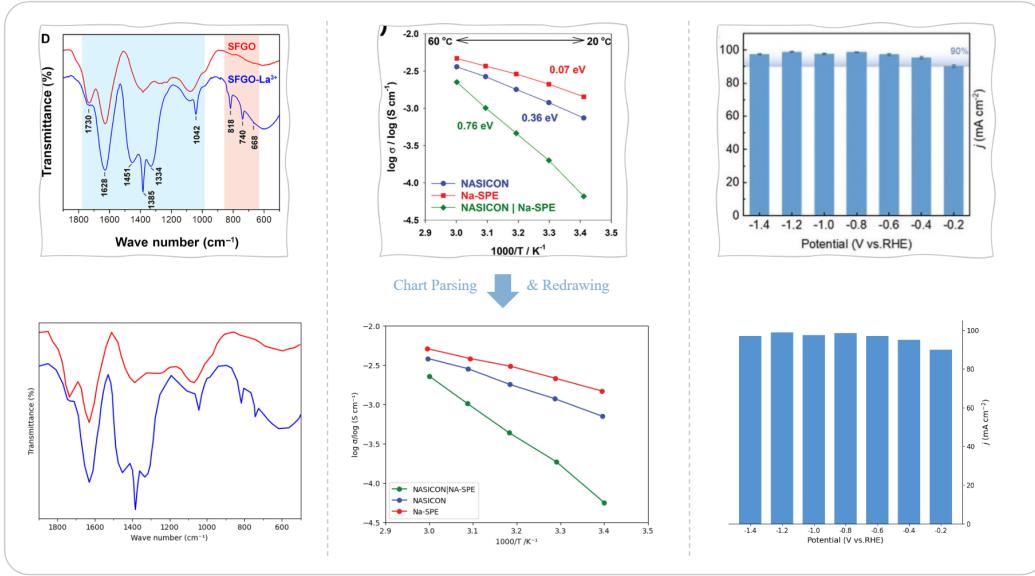


Figure 6: Examples of chart recognition results produced by Uni-Parser. Chart images are first parsed into underlying data tables, which are then re-rendered in a style consistent with the original figures for visualization and presentation.

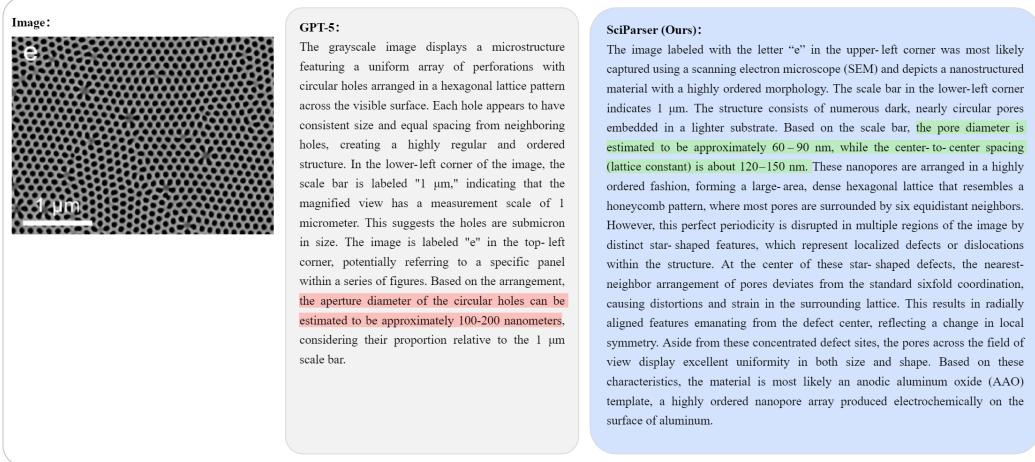


Figure 7: Examples of scientific figure captioning results produced by Uni-Parser. SciParser, a dedicated scientific image captioning submodule within Uni-Parser, converts scientific figures into high-information-density textual descriptions that capture key semantic and structural attributes, facilitating downstream understanding, retrieval, and database construction.

2.8 Reading Order Recovery

To accurately reconstruct the reading order of content blocks on each page, our system integrates a set of spatial and semantic heuristics tailored for complex, real-world documents.

XY-cut. This method recursively partitions the page along dominant whitespace regions, yielding a hierarchical binary reading tree that captures coarse layout structure.

Gap-tree analysis. It leverages inter-block whitespace, geometric proximity, and alignment cues to infer plausible reading flows, particularly in dense or irregular layouts.

Group-based strategies. These strategies cluster semantically related elements—even when spatially distant—such as linking figures with captions or molecules with identifiers, thereby preserving logical semantics prior to global ordering.

By combining these complementary techniques, the system produces a natural and coherent reading sequence across multi-column, multi-lingual, and multi-modal layouts.

2.9 Cross-column and Cross-page Consolidation

After establishing the per-page reading order, the system further consolidates items that are visually separated but semantically continuous:

- **Cross-column merging:** reconnecting paragraph fragments split by column boundaries using text-flow continuity and linguistic coherence.
- **Cross-page merging:** linking entities that span pages—such as long tables, multi-step reaction schemes, or running paragraphs—based on layout carry-over cues and semantic consistency.
- **Multi-modal linkage:** associating diagrams with their identifiers or descriptions even when they appear on adjacent pages.

This second-stage consolidation restores logically unified content into coherent units, improving the fidelity of the reconstructed document structure.

3 Infrastructure

3.1 Distributed Multi-Expert Architecture

Microservice architecture: Uni-Parser adopts a microservice-based multi-expert architecture that enables large-scale distributed inference. Each modality-specific expert (e.g., text, molecules, formulas, tables, reactions, and charts) is deployed as an independent microservice, with multiple nodes running in parallel to process inference requests. Layout analysis first partitions documents into batches, which are distributed across nodes for concurrent processing. Detected regions are then enqueued into modality-specific task queues, where batched requests are asynchronously dispatched to the corresponding expert services. Finally, outputs from all modules are aggregated into a unified generalized OCR representation, followed by post-processing and structured result generation.

Dynamic load balancing: A fine-grained scheduling layer dynamically allocates computational resources both within and across modules. This design supports adaptive scaling under varying workloads, prevents bottlenecks in individual experts, and ensures stable throughput in heterogeneous multi-modal parsing.

Pipeline Parallel: The inference runtime is optimized for efficient GPU parallelism and scheduling, employing multi-process server execution, micro-batching, and asynchronous task management to minimize idle time and sustain high GPU occupancy, while simultaneously ensuring effective utilization of CPU and memory resources across heterogeneous workloads. In particular, CPU pre-processing and post-processing, GPU model inference, and inter-service data transfer are orchestrated in a pipelined manner, enabling time-overlapping execution across stages and thereby further reducing latency in each instance. An analysis of bubble time is presented in Figure 8.

Decoupled module updates: Independent component upgrades that support rapid iteration and performance tuning without full system redeployment, with zero service interruption during updates.

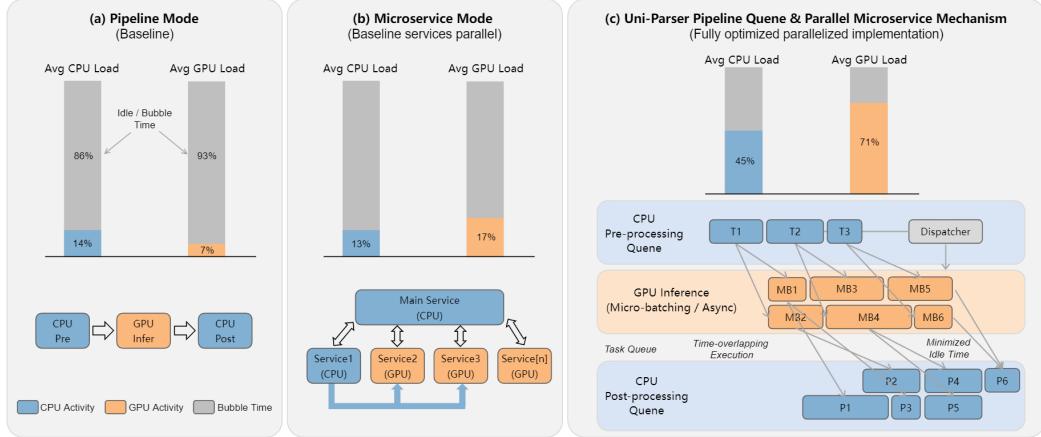


Figure 8: Comparison of document parsing infrastructures. (a) Sequential pipeline: Most existing document parsing frameworks adopt a strictly serial workflow, in which layout analysis is completed first, followed by generalized OCR tasks and subsequent post-processing. (b) Microservice-based parallelism: The tasks are dispatched to multiple microservices that operate in parallel, and the final results are aggregated through a gather stage. (c) Our Uni-Parser pipeline-parallel design: By enabling more frequent lightweight communication and non-blocking execution across modules, Uni-Parser fully exploits parallelism across heterogeneous microservices. Computation and communication are effectively overlapped across GPU and CPU resources, substantially improving throughput and reducing end-to-end latency.

This enables a self-reinforcing data flywheel loop, allowing models to be updated on an hourly basis and continuously enhanced with newly acquired active learning samples.

3.2 Deployment Scaling

Uni-Parser is designed for elastic scaling in distributed inference environments. System throughput can be increased almost linearly by expanding the number of nodes assigned to each load-balanced microservice. Benchmark experiments demonstrate that the average parsing speed per PDF page scales almost linearly with the number of backend GPUs.

The Uni-Parser system has been scaled to a cluster of 240 NVIDIA L40 (48GB) GPUs, with each GPU allocated 22 CPU cores and 90GB of memory. Using this configuration, the system parses more than 16 million documents in less than 6 days (in Uni-Parser fast mode). These results demonstrate that the Uni-Parser infrastructure is highly efficient, economically scalable, and robustly stable under large-scale processing workloads.

4 Data Engineering

4.1 Uni-Parser Data Engine

The *Uni-Parser Data Engine* is designed to efficiently generate and curate high-quality training data for PDF parsing models, combining synthetic generation, active learning on real documents, and self-training on unlabeled data to minimize human annotation effort while maintaining high accuracy.

Step 1: Synthetic Data for Model Bootstrapping. Since all PDFs are human-generated rather than naturally structured, we employ large-scale synthetic data to bootstrap various models, including layout detection and several generalized OCR models. Diverse synthetic documents are generated and augmented extensively to pretrain models, providing a strong initialization that accelerates downstream learning and improves generalization.

Step 2: Active Learning on Real Documents. After the model is equipped with basic capabilities using synthetic data, we leverage a large corpus of real-world documents to iteratively enhance its performance through active learning. Our in-house training datasets are curated from:

- 170 million pages of scientific literature spanning multiple disciplines;
- 140 million pages of patent documents from patent offices worldwide;
- 20 million pages of books, reports, and other documents sourced from Fine-PDFs [30].

An active learning data flywheel, similar to the one used in MolParser [23], is employed to select the most informative samples for annotation. This is combined with a curriculum learning strategy that progressively trains the models on increasingly complex examples.

Step 3: Self-Training on Unlabeled Data. To further scale training without additional manual labeling, we apply self-training on large collections of unlabeled documents. Predictions from multiple ensembled models are aggregated to estimate confidence scores, and only high-confidence predictions are treated as pseudo-labels. This approach effectively expands the training set at minimal cost while maintaining label quality.

By combining synthetic pretraining, active learning, and self-training, we construct a data flywheel that substantially reduces the amount of human annotations required, accelerates model convergence, and enables high-throughput, low-cost data preparation. Using this three-step approach, we reduce the total annotation volume by 95% and cut per-page or per-block annotation time by 90%, with 90% of annotations for our various models completed in just two months, ensuring rapid, scalable, and efficient model development for large-scale PDF parsing tasks.

4.2 Uni-Miner Annotation Platform

To better establish a data flywheel for Uni-Parser with an effective human-in-the-loop pipeline, we designed the *Uni-Miner Annotation Platform* to ensure that samples selected through active learning can be corrected and refined by human annotators with high quality and low cost. The platform integrates a molecular drawing interface and supports annotation across multiple modalities, including molecular structures, text, and layout elements. In addition, we incorporate a recommendation module that assigns the most suitable samples to domain experts based on their specialization and past performance. Each annotator and reviewer is dynamically scored according to accuracy metrics, enabling the system to route the most critical or ambiguous cases to the most experienced and reliable individuals. This design significantly reduces annotation cost, improves throughput, and ensures consistent data quality through mechanisms such as inter-annotator agreement monitoring. In practice, each sample is reviewed by at least two annotators, providing further quality assurance for the downstream learning process.

5 Performance

5.1 Construction of Uni-Parser Benchmark

Uni-Parser benchmark dataset targets layout detection and semantic contents recognition in scientific documents, comprising 150 PDFs spanning international patents and research articles. The selection emphasizes diversity in structure, subject, domain, and language. The distribution of this benchmark is shown in Table 2.

Table 2: Distribution of document sources in Uni-Parser benchmark.

Source Type	Sub-Source	#PDFs	Total Pages
Patent Documents	From 20 Patent Offices (5 langs)	50	1455
	xRxiv (bioRxiv, medRxiv, ChemRxiv, etc.)	35	696
	arxiv	20	231
Scientific Articles	ChinaXiv	15	168
	Nature Communications	15	164
	Scientific Reports	15	173
Total (All Documents)		150	2887

6 Applications

Uni-Parser opens up a wide range of downstream applications, spanning literature understanding, structured knowledge extraction, patent analysis, and large-scale data generation. Together, these applications demonstrate the framework’s versatility in advancing both scientific research and industrial practice.

Document Understanding. Uni-Parser significantly enhances scientific literature workflows by supporting automatic document summarization, document-level question answering [38], paper-to-poster [39] and paper-to-PPT [40] generation, as well as intelligent retrieval and deep research [41]. Collectively, these capabilities streamline both knowledge consumption and dissemination for researchers across diverse disciplines.

Structured Data Extraction. By converting unstructured documents into structured representations, Uni-Parser supports the large-scale construction of domain-specific databases, such as paper citation database, scholar database, molecular libraries, reaction repositories, bioactivity database [31], experimental characterization database [29, 42], and comprehensive entity knowledge bases. Such resources are critical for accelerating data-driven scientific discovery.

Patent Retrieval and Protection. The framework further facilitates patent retrieval and prior-art verification, providing robust support for innovation discovery and intellectual property protection [43, 44]. This enables more efficient navigation of complex patent landscapes in scientific and industrial settings.

Foundation Model Training. Finally, Uni-Parser can serve as a powerful engine for large-scale training data generation, supplying high-quality structured inputs that reduce the cost and effort of manual curation. This capability is particularly valuable for advancing foundation models in scientific domains, where large, reliable datasets are indispensable [45–47].

7 Failed Approaches

We summarize things that didn’t work during the development of Uni-Parser.

In most scenarios of document intelligence, large-scale pretraining on synthetic data is a simple and effective strategy, because natural documents are largely computer-rendered and therefore exhibit strong domain consistency. However, as discussed in Section 2.2, this intuition breaks down for layout recognition in scientific literature. The layouts of scientific articles are shaped extensively by human editorial practices and creative design choices, making them difficult for heuristic layout engines to reproduce. As a result, synthetic data not only fails to cover the diversity of real-world layouts but can even distort the true data distribution.

For the OCSR task, atom–bond (graph-based) methods are intuitively appealing and have a long history of successful applications; they were also the first direction we explored. While these methods offer clear advantages for handling chirality, they struggle with the wide variety of challenging cases present in real scientific literature. Their strong reliance on rigid, hand-crafted rules fundamentally limits scalability—simply increasing training data provides little benefit. In addition, these methods require substantially more manual annotation effort, typically over 20 \times that of end-to-end approaches, further constraining their practicality. As a result, compared with end-to-end models, graph-based methods suffer from lower performance ceilings, slower inference, and prohibitively high annotation costs.

8 Future Work

Enhancing Core Components. Uni-Parser’s distributed and modular pipeline architecture allows individual components to be easily upgraded or replaced, facilitating continuous improvement across different document types. We plan to iteratively update the core components to further improve extraction quality across diverse document types:

- **Layout detection:** Our current models are primarily tailored to scientific and patent documents. However, the diversity of document types and layouts is virtually limitless. We will continue

to enhance our layout recognition models to support an increasingly broader range of scenarios, including newspapers and magazines, PPT slides, various book formats, and financial statements.

- **OCSR model:** Although our MolParser 1.5 already outperforms previous state-of-the-art methods, the recognition of chiral molecules still presents significant challenges. We will focus on exploring how to address the challenge of chirality recognition within end-to-end OCSR models.
- **Chemical reaction understanding:** Parsing chemical reactions in real-world literature remains highly challenging, with substantial room for improving generalization performance.
- **Chart understanding:** Currently, all existing Chart2Table models and general-purpose MLLMs fall far short of meeting industrial-level requirements for parsing charts in scientific literature, which exhibit a wide variety of types and styles. Chart parsing therefore still holds substantial room for further exploration.
- **Reading order:** We plan to incorporate machine learning-based reading order predictors, to enhance the generalization ability of reading order prediction under complex layouts.
- **Deployment optimization:** Techniques such as quantization (PTQ and QAT), distillation, pruning, and other inference acceleration methods will be explored, along with support for diverse hardware platforms, including Ascend NPUs.

Uni-Parser-Tools for Easy Access. We will release a fully packaged, open-source toolkit, *Uni-Parser-Tools*, which enables remote access to Uni-Parser without requiring local computational resources. This toolkit will include example pipelines for downstream applications, allowing users to quickly leverage Uni-Parser to construct structured scientific databases and unlock generative AI applications for scientific discovery.

Benchmark Construction. Existing benchmarks, such as OmniDocBench and our Uni-Parser benchmark, are constrained by inconsistencies in layout recognition and the heterogeneous output formats of generalized OCR modules. Moreover, they struggle to adequately represent complex layouts and cross-page content. We aim to investigate more robust approaches to benchmark construction, particularly task-driven benchmarks tailored to downstream applications, including document understanding and structured data extraction. Such efforts will enable more fair, comprehensive, and informative evaluations of parsing performance.

References

- [1] N. Livathinos, C. Auer, M. Lysak, A. Nassar, M. Dolfi, P. Vagenas, C. B. Ramis, M. Omenetti, K. Dinkla, Y. Kim *et al.*, “Docling: An efficient open-source toolkit for ai-driven document conversion,” *arXiv preprint arXiv:2501.17887*, 2025.
- [2] C. Cui, T. Sun, M. Lin, T. Gao, Y. Zhang, J. Liu, X. Wang, Z. Zhang, C. Zhou, H. Liu *et al.*, “Paddleocr 3.0 technical report,” *arXiv preprint arXiv:2507.05595*, 2025.
- [3] B. Wang, C. Xu, X. Zhao, L. Ouyang, F. Wu, Z. Zhao, R. Xu, K. Liu, Y. Qu, F. Shang *et al.*, “Mineru: An open-source solution for precise document content extraction,” *arXiv preprint arXiv:2409.18839*, 2024.
- [4] MathPix Team, “MathPix,” <https://mathpix.com/>, 2025.
- [5] L. Blecher, G. Cucurull, T. Scialom, and R. Stojnic, “Nougat: Neural optical understanding for academic documents,” *arXiv preprint arXiv:2308.13418*, 2023.
- [6] H. Wei, C. Liu, J. Chen, J. Wang, L. Kong, Y. Xu, Z. Ge, L. Zhao, J. Sun, Y. Peng *et al.*, “General ocr theory: Towards ocr-2.0 via a unified end-to-end model,” *arXiv preprint arXiv:2409.01704*, 2024.
- [7] A. Nassar, A. Marafioti, M. Omenetti, M. Lysak, N. Livathinos, C. Auer, L. Morin, R. T. de Lima, Y. Kim, A. S. Gurbuz *et al.*, “Smoldocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion,” *arXiv preprint arXiv:2503.11576*, 2025.
- [8] Z. Li, Y. Liu, Q. Liu, Z. Ma, Z. Zhang, S. Zhang, Z. Guo, J. Zhang, X. Wang, and X. Bai, “Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm,” *arXiv preprint arXiv:2506.05218*, 2025.
- [9] OCRFlux Team, “OCRFlux: Mastering complex layouts and seamless page merging,” <https://ocrflux.pdfparser.io/>, 2025.

- [10] rednote-hilab, “dots.ocr: Multilingual document layout parsing in a single vision-language model,” <https://huggingface.co/rednote-hilab/dots.ocr>, 2025.
- [11] Mistral Team, “MistralOCR,” <https://mistral.ai/news/>, 2025.
- [12] Z. Zhao, H. Kang, B. Wang, and C. He, “Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception,” *arXiv preprint arXiv:2410.12628*, 2024.
- [13] B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, and P. Staar, “Doclaynet: A large human-annotated dataset for document-layout segmentation,” in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022, pp. 3743–3751.
- [14] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, “Detrs beat yolos on real-time object detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 16965–16974.
- [15] Y. Tian, Q. Ye, and D. Doermann, “Yolov12: Attention-centric real-time object detectors,” *arXiv preprint arXiv:2502.12524*, 2025.
- [16] Y. Peng, H. Li, P. Wu, Y. Zhang, X. Sun, and F. Wu, “D-fine: Redefine regression task in detrs as fine-grained distribution refinement,” *arXiv preprint arXiv:2410.13842*, 2024.
- [17] G. Jocher, J. Qiu, and A. Chaurasia, “Ultralytics yolo,” Ultralytics, 2025. [Online]. Available: <https://ultralytics.com>
- [18] A. Steiner, A. S. Pinto, M. Tschannen, D. Keysers, X. Wang, Y. Bitton, A. Gritsenko, M. Minderer, A. Sherbondy, S. Long *et al.*, “Paligemma 2: A family of versatile vlms for transfer,” *arXiv preprint arXiv:2412.03555*, 2024.
- [19] C. Cui, T. Sun, S. Liang, T. Gao, Z. Zhang, J. Liu, X. Wang, C. Zhou, H. Liu, M. Lin, Y. Zhang, Y. Zhang, H. Zheng, J. Zhang, J. Zhang, Y. Liu, D. Yu, and Y. Ma, “Paddleocr-vl: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model,” 2025. [Online]. Available: <https://arxiv.org/abs/2510.14528>
- [20] C. Li, R. Guo, J. Zhou, M. An, Y. Du, L. Zhu, Y. Liu, X. Hu, and D. Yu, “Pp-structurev2: A stronger document analysis system,” *arXiv preprint arXiv:2210.05391*, 2022.
- [21] X. Zhong, E. ShafieiBavani, and A. J. Yipes, “Image-based table recognition: data, model, and evaluation,” *arXiv preprint arXiv:1911.10683*, 2019.
- [22] A. Nassar, N. Livathinos, M. Lysak, and P. Staar, “Tableformer: Table structure understanding with transformers,” *arXiv preprint arXiv:2203.01017*, 2022.
- [23] X. Fang, J. Wang, X. Cai, S. Chen, S. Yang, H. Tao, N. Wang, L. Yao, L. Zhang, and G. Ke, “Molparser: End-to-end visual recognition of molecule structures in the wild,” *arXiv preprint arXiv:2411.11098*, 2024.
- [24] Y. Qian, J. Guo, Z. Tu, Z. Li, C. W. Coley, and R. Barzilay, “Molscribe: robust molecular structure recognition with image-to-graph generation,” *Journal of Chemical Information and Modeling*, vol. 63, no. 7, pp. 1925–1934, 2023.
- [25] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, “Qwen2. 5-vl technical report,” *arXiv preprint arXiv:2502.13923*, 2025.
- [26] OpenAI, “Introducing gpt-5,” 2025. [Online]. Available: <https://openai.com/index/introducing-gpt-5/>
- [27] DeepMind, “Gemini 2.5 pro,” 2025. [Online]. Available: <https://deepmind.google/models/gemini/pro/>
- [28] A. Mirza, N. Alampara, M. Ríos-García, M. Abdelalim, J. Butler, B. Connolly, T. Dogan, M. Nezhurina, B. Şen, S. Tirunagari *et al.*, “Chempile: A 250gb diverse and curated dataset for chemical foundation models,” *arXiv preprint arXiv:2505.12534*, 2025.
- [29] Z. Xia, Y. Li, S. Tang, Z. Fan, X. Fang, H. Tao, X. Cai, G. Ke, L. Zhang, Y. Hong *et al.*, “Uniem-3m: A universal electron micrograph dataset for microstructural segmentation and generation,” *arXiv preprint arXiv:2508.16239*, 2025.
- [30] H. Kydlíček, G. Penedo, and L. von Werra, “Finepdfs,” <https://huggingface.co/datasets/HuggingFaceFW/finepdfs>, 2025.

- [31] J. Yan, J. Zhu, Y. Yang, Q. Liu, K. Zhang, Z. Zhang, X. Liu, B. Zhang, K. Gao, J. Xiao *et al.*, “Biominer: A multi-modal system for automated mining of protein-ligand bioactivity data from literature,” *bioRxiv*, 2025.
- [32] XtalPi, “Molvision,” 2025. [Online]. Available: <https://www.xinsight-ai.com/mol-recognition?mol-recognition=%2Fmol-recognition>
- [33] J. Xiong, X. Liu, Z. Li, H. Xiao, G. Wang, Z. Niu, C. Fei, F. Zhong, G. Wang, W. Zhang *et al.*, “ α extractor: a system for automatic extraction of chemical information from biomedical literature,” *Sci. China: Life Sci.*, vol. 67, pp. 618–621, 2023.
- [34] Y. Xu, J. Xiao, C.-H. Chou, J. Zhang, J. Zhu, Q. Hu, H. Li, N. Han, B. Liu, S. Zhang, J. Han, Z. Zhang, S. Zhang, W. Zhang, L. Lai, and J. Pei, “Molminer: You only look once for chemical structure recognition,” *Journal of Chemical Information and Modeling*, vol. 62, no. 22, pp. 5321–5328, 2022.
- [35] Y. Chen, C. T. Leung, Y. Huang, J. Sun, H. Chen, and H. Gao, “Molnextr: A generalized deep learning model for molecular image recognition,” *arXiv preprint arXiv:2403.03691*, 2024.
- [36] K. Rajan, H. O. Brinkhaus, M. I. Agea, A. Zielesny, and C. Steinbeck, “Decimer. ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications,” *Nature communications*, vol. 14, no. 1, p. 5045, 2023.
- [37] MinerU Team, “Mineru.chem,” 2025. [Online]. Available: <https://mineru.net/chem/>
- [38] H. Cai, X. Cai, S. Yang, J. Wang, L. Yao, Z. Gao, J. Chang, S. Li, M. Xu, C. Wang *et al.*, “Uni-smart: universal science multimodal analysis and research transformer,” *arXiv preprint arXiv:2403.10301*, 2024.
- [39] W. Pang, K. Q. Lin, X. Jian, X. He, and P. Torr, “Paper2poster: Towards multimodal poster automation from scientific papers,” *arXiv preprint arXiv:2505.21497*, 2025.
- [40] D.-A. D. Team, “Dataflow-agent: Ai-powered data & paper workflow orchestration platform,” <https://github.com/OpenDCAI/DataFlow-Agent>, 2025.
- [41] OpenAI, “Deep research,” <https://openai.com/deep-research>, 2025, [AI research tool].
- [42] F. Xu, W. Guo, F. Wang, L. Yao, H. Wang, F. Tang, Z. Gao, L. Zhang, W. E, Z.-Q. Tian *et al.*, “Toward a unified benchmark and framework for deep learning-based prediction of nuclear magnetic resonance chemical shifts,” *Nature Computational Science*, pp. 1–9, 2025.
- [43] Y. Shi, S. Li, T. Zhang, X. Fang, J. Wang, Z. Liu, G. Zhao, Z. Zhu, Z. Gao, R. Zhong *et al.*, “Intelligent system for automated molecular patent infringement assessment,” *arXiv preprint arXiv:2412.07819*, 2024.
- [44] J. Zhuang, K. Li, J. Hou, M. Xu, Z. Gao, and H. Cai, “Doc2sar: A synergistic framework for high-fidelity extraction of structure-activity relationships from scientific documents,” *arXiv preprint arXiv:2506.21625*, 2025.
- [45] S. Li, J. Huang, J. Zhuang, Y. Shi, X. Cai, M. Xu, X. Wang, L. Zhang, G. Ke, and H. Cai, “Scilitllm: How to adapt llms for scientific literature understanding,” *arXiv preprint arXiv:2408.15545*, 2024.
- [46] N. Liao, X. Wang, Z. Lin, W. Guo, F. Hong, S. Song, G. Yu, Z. Zhao, S. Xie, L. Wei *et al.*, “Innovator: Scientific continued pretraining with fine-grained moe upcycling,” *arXiv preprint arXiv:2507.18671*, 2025.
- [47] D. D. Team, “Dataflow: A unified framework for data-centric ai,” <https://github.com/OpenDCAI/DataFlow>, 2025, accessed: 2025-07-08.

A Authorship and Acknowledgments

Please cite this work as “DP Technology (2025)”.

Correspondence regarding this technical report can be sent to fangxi@dp.tech

Principal Contributor
Haoyi Tao

Core Contributors*
Chaozheng Huang
Han Lyu
Haocheng Lu
Shuwen Yang
Suyang Zhong
Xi Fang

Project Lead
Xi Fang

Program Leads*
Guolin Ke
Linfeng Zhang
Xinyu Li

Contributors & Acknowledgments*
Changhong Chen
Chenkai Wu
Daoqin Cui
Fanjie Xu
Hang Zheng
Hanzheng Li
Hengxin Cai
Jiale Ren
Jialu Shen
Jiankun Wang
Jiaxi Zhuang
Jindi Guo
Jingwen Deng
Lin Yao
Mingjun Xu
Nan Wang
Ning Wang
Qingguo Zhou
Shangqian Chen
Shaojie Chen
Shengyu Li
Sian Chen
Xiaochen Cai
Xiaohong Ji
Xinyu Xiong
Xuan Xie
Yanhui Hong
Yaorui Shi
Yaqi Li
Yixuan Li
Yuan Gao
Zhenhao Wong
Zhifeng Gao
Zhiyue Wang

*Contributors listed in alphabetized order.

Raw Image	Markush				
	MinerU.Chem	MolScribe	MolVision	AlphaExtractor	MolParser 1.5
Connection Point					
Raw Image	MinerU.Chem	MolScribe	MolVision	AlphaExtractor	MolParser 1.5
Colored					
Raw Image	MinerU.Chem	MolScribe	MolVision	AlphaExtractor	MolParser 1.5

Raw Image	Noise				
	MinerU.Chem	MolScribe	MolVision	AlphaExtractor	MolParser 1.5

B.2 Uni-Parser Qualitative Examples

Examples here.