Datasets for Scientific Literature Understanding: A Survey

Anonymous ACL submission

Abstract

Empowering machines to understand scientific literature is crucial for accelerating scientific discovery and advancing the AI for Science (AI4S) paradigm. In this paper, we present a comprehensive survey of datasets serving this domain. We propose a systematic taxonomy that organizes resources spanning structural parsing, text understanding, multimodal understanding and pre-training/instruction fine-tuning. Beyond a structured overview, we discuss the evolution of the field, elucidating how the emergence of Large Language Models (LLMs) has reshaped research priorities of dataset construction. By synthesizing existing datasets and identifying critical future directions, this work provides a roadmap for advancing intelligent scientific research systems.

1 Introduction

Scientific literature embodies core domain knowledge and constitutes the primary vehicle for disseminating scientific discoveries (Hanson et al., 2024; Larsen and Von Ins, 2010). With the unprecedented proliferation of scholarly publications, researchers increasingly face challenges in comprehensively understanding and reasoning over the expanding corpus of knowledge (Landhuis, 2016). Consequently, empowering machines to comprehend scientific literature and advancing toward complex reasoning has become a critical research direction. This capability not only enables researchers to efficiently acquire and synthesize novel insights, accelerating scientific discovery, but also underpins the development of the AI for Science (AI4S) paradigm (Xu et al., 2021), allowing intelligent systems to assist in complex scientific analyses and foster innovation. Concurrently, the rapid advancement of Large Language Models (LLMs) (Zhao et al., 2023; Naveed et al., 2025; Minaee et al., 2024) and multimodal learning techniques (Wu et al., 2023; Zhang et al., 2024b; Yin et al., 2024) present transformative opportunities while posing novel methodological and computational challenges. Therefore, developing effective methodologies for scientific literature understanding constitutes both a practical imperative and a strategic frontier for AI-driven scientific research.

This paper presents a comprehensive overview of datasets for scientific literature understanding. To capture the breadth of available resources, we categorize datasets into four primary groups. As illustrated in Figure 1, the first three categories map the transition from low-level visual perception to high-level multimodal reasoning, while the fourth category provides the essential infrastructure for building and aligning the models. Details of the datasets are provided in Appendix A.

- **Structural Parsing** establishes the perceptual foundation by transforming complex document layouts into machine-interpretable representations.
- **Text Understanding** advances to the semantic level, involving extracting structured knowledge, summaries, or answering questions based on textual content.
- **Multimodal Understanding** emphasizes the integration of heterogeneous modalities to enable domain-specific, evidence-based reasoning.
- **Pre-training and Instruction Fine-tuning** provides LLMs the large-scale corpora for domain knowledge acquisition and the alignment data required for accurate, contextually grounded outputs.

To the best of our knowledge, this work presents the first comprehensive survey in this rapidly evolving domain. Through this structured

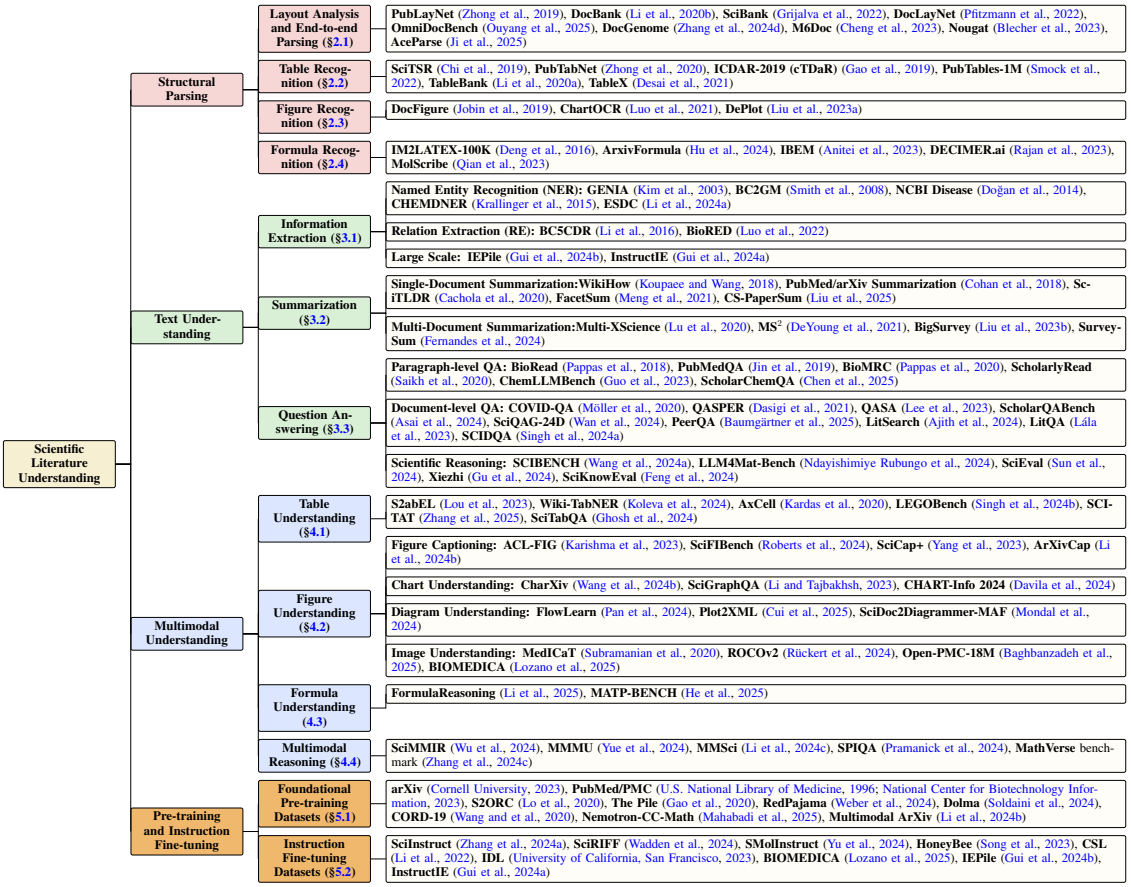

Figure 1: Taxonomy of Datasets for Scientific Literature Understanding

overview, we aim to equip researchers with a holistic view of the dataset landscape, facilitating informed selection and application in downstream tasks. Moreover, this survey reflects the field's evolutionary trajectory and shifting priorities, particularly highlighting how the emergence of LLMs has reshaped the requirements for data scale, quality, and multimodal integration. By synthesizing the existing state of datasets and identifying key future directions, this survey serves as a foundational reference for advancing scientific literature understanding, which remains a core prerequisite for the successful realization of the AI4S vision.

2 Structural Parsing Datasets

Scientific literature exhibits complex document layouts that tightly integrate text, tables, figures, and formulas. Structural parsing aims to transform these heterogeneous visual signals into machine-interpretable representations, forming a perceptual foundation for higher-level scientific document understanding and reasoning.

2.1 Layout Analysis and End-to-end Parsing

Layout analysis converts the visual geometry of scientific documents into hierarchical structures, while Optical Character Recognition (OCR) transforms visual text into machine-readable content. Gemelli et al. (2024) has already provided a thorough overview of this field. Early large-scale datasets such as **PubLayNet** (Zhong et al., 2019) and **DocBank** (Li et al., 2020b) align Portable Document Format (PDF) elements with markup to support layout detection and reading-order reconstruction. **SciBank** (Grijalva et al., 2022) complements existing benchmarks by providing a human-validated dataset of over 74K scientific pages with comprehensive annotations across 12 distinct layout regions. **DocLayNet** (Pfitzmann et al., 2022) and multilingual datasets like **OmniDocBench** (Ouyang et al., 2025) extend this to complex and non-English layouts. **DocGenome** (Zhang et al., 2024d) provides a large-scale, structured corpus of 500K arXiv papers with fine-grained layout, LaTeX, and logical annotations, serving as a comprehensive benchmark for multi-modal scientific document analysis. For historical and scanned ma-

terials, **M6Doc** (Cheng et al., 2023) improves robustness across document types. More recent end-to-end frameworks like **Nougat** (Blecher et al., 2023) and **AceParse** (Ji et al., 2025) integrate layout parsing, OCR, and table/formula extraction for comprehensive PDF understanding.

Dataset development has progressed from single-page layout annotation to multilingual and end-to-end structural corpora. Key challenges include establishing fine-grained element correspondence (e.g., spanning cells), recovering consistent reading orders across heterogeneous formats, and ensuring coverage for low-resource languages and historical documents. Future datasets should prioritize structural completeness and cross-domain generalization.

2.2 Table Recognition

Tables condense scientific evidence, requiring robust structural interpretation for automated parsing. Foundational datasets such as **SciTSR** (Chi et al., 2019) and **PubTabNet** (Zhong et al., 2020) establishes datasets for parsing table structures. The **ICDAR-2019 (cTDaR)** competition (Gao et al., 2019) further standardizes this task, focusing on complex table layouts. Large-scale datasets like **PubTables-1M** (Smock et al., 2022), **TableBank** (Li et al., 2020a) and **TableX** (Desai et al., 2021) provide extensive data sourced from scientific articles and diverse document formats, supporting the development of generalized models for table structural recovery.

Despite these advancements, existing datasets remain plagued by alignment noise from heuristic pipelines and struggle with non-standard topologies such as borderless or nested tables. Future efforts must therefore prioritize precise structural annotations and unified evaluation protocols to ensure robust performance on complex layouts.

2.3 Figure Recognition

Figure recognition focuses on categorizing visual content and resolving internal structures. **DocFigure** (Jobin et al., 2019) establishes a foundational benchmark for fine-grained figure classification, distinguishing among 28 distinct categories such as heatmaps and box plots. Moving towards structural recovery, **ChartOCR** (Luo et al., 2021) utilizes a deep keypoint detection framework to extract raw data values directly from chart images. **DePlot** (Liu et al., 2023a) reformulates the recognition task as a visual-to-language problem, training models to translate chart images into linearized textual tables for direct downstream processing.

Dataset development is shifting focus from simple categorization to fine-grained element parsing and data recovery. However, significant challenges remain in handling real-world scientific illustrations, particularly in disentangling multi-panel compound figures and resolving the precise alignment between legends, axes, and data points. Future benchmarks should ensure models can accurately translate visual signals into numerical evidence within the broader document context.

2.4 Formula Recognition

Formula recognition bridges the gap between visual 2D layouts and semantic sequences. Datasets such as **IM2LATEX-100K** (Deng et al., 2016) and **ArxivFormula** (Hu et al., 2024) facilitate image-to-markup learning. Layout-aware datasets like **IBEM** (Anitei et al., 2023) contextualize equations within document structure. In applied sciences, resources like **DECIMER.ai** (Rajan et al., 2023) and **MolScribe** (Qian et al., 2023) extend parsing to chemical and molecular formulas.

The evolution of formula datasets reflects a shift from isolated symbol recognition to the contextualization of mathematical expressions within document layouts. Challenges remain in cross-references with text and tables, and maintaining consistency across symbolic, visual, and contextual information to support higher-level reasoning. Future datasets should prioritize high-resolution structural annotations and robust handling of inline formulas, ensuring that parsing models can distinguish and accurately reconstruct dense scientific notations amidst textual content.

3 Text Understanding Datasets

Text remains the predominant modality in scientific literature, serving as the principal medium of claims, methods, and evidence. In this section, we review representative datasets for three major tasks: information extraction, summarization, and question answering, restricting our scope to purely text-based tasks.

3.1 Information Extraction

Scientific Information Extraction (IE) seeks to convert unstructured scholarly text into structured representations by identifying entities, relations, and complex semantic structures.

Early Named Entity Recognition (NER) datasets establish the foundation for structured text understanding. **GENIA** (Kim et al., 2003) pioneers large-scale annotation of biomedical abstracts, enabling the transition from rule-based systems to statistical learning. **BC2GM** (Smith et al., 2008) specializes in gene mentions, while **NCBI Disease** (Doğan et al., 2014) links disease mentions to MeSH and OMIM vocabularies. **CHEMDNER** (Krallinger et al., 2015) provides over 80K annotated chemical entities, establishing a dataset for chemical IE. Beyond biomedicine, **ESDC** (Li et al., 2024a) extends NER to geosciences, annotating geographic entities and observational indicators.

Scientific Relation Extraction (RE) emerges to model semantic connections among entities. **BC5CDR** (Li et al., 2016) jointly evaluates NER and chemicaldisease relations, while **BioRED** (Luo et al., 2022) scales to document-level RE with over 6,500 relations across six entity types, distinguishing novel findings from background knowledge.

Large scale IE datasets are created for LLMs: **IEPile** (Gui et al., 2024b) integrates 33 datasets across multiple domains into a bilingual corpus (0.32B characters) with instruction augmentation and hard-negative sampling; **InstructIE** (Gui et al., 2024a) provides a bilingual instruction dataset covering 12 themes from Wikipedia and Wikidata to guide triple extraction via natural-language prompts.

Scientific IE datasets have evolved from entity-level annotation to relation-aware and instruction-aligned corpora. Core challenges include domain coverage, schema diversity, and integration of multilingual or cross-domain knowledge. Future datasets should emphasize scalability, annotation consistency, and alignment with reasoning tasks, enabling robust extraction of structured scientific knowledge beyond superficial labeling.

3.2 Summarization

Summarization condenses lengthy, discourse-rich documents into concise, faithful representations, capturing contributions, methods, and findings. Datasets are categorized into single-document summarization (SDS) and multi-document summarization (MDS).

Single-Document Summarization Datasets such as **WikiHow** (Koupaee and Wang, 2018) provide 230K instructional articles with sentence- and paragraph-level summaries. **PubMed/arXiv Summarization** (Cohan et al., 2018) align full papers with abstracts (215K from arXiv, 133K from PubMed) for discourse-aware modeling. **SciTLDR** (Cachola et al., 2020) offers extreme-compression summaries for more than 5,400 papers, while **FacetSum** (Meng et al., 2021) structures 60K Emerald journal articles into purpose, method, findings, and value facets. In the LLM era, **CS-PaperSum** (Liu et al., 2025) scales SDS to 92K computer science papers with AI-generated structured summaries.

Multi-Document Summarization **Multi-XScience** (Lu et al., 2020) generates related-work sections from abstracts and citations (30K examples). **MS^2** (DeYoung et al., 2021) links 20K biomedical reviews to 470K abstracts, incorporating structured PICO elements. **BigSurvey** (Liu et al., 2023b) covers 7,000 surveys and 434K referenced abstracts with section-structured synthesis. **SurveySum** (Fernandes et al., 2024) aligns full texts of cited papers with survey sections, enabling LLM-based retrieval-augmented pipelines.

Scientific summarization spans SDS and MDS, and some datasets also adopts facet-structured outputs to support more fine-grained controllability. Key considerations include capturing discourse structures, aligning summaries with evidence, and supporting multi-granularity output. Future datasets should provide structured annotations that enable controllable, knowledge-centric summarization, while accommodating long-context reasoning and cross-document integration.

3.3 Question Answering

Question Answering (QA) evaluates models ability to retrieve, integrate, and reason over textual evidence. Datasets in this section focus on reasoning over pure text (paragraphs, abstracts, or full-text without visual or tabular grounding).

Paragraph-level QA Early datasets such as **BioRead** (Pappas et al., 2018) and **PubMedQA** (Jin et al., 2019) focus on span extraction or categorical (yes/no/maybe) answers from abstracts. **BioMRC** (Pappas et al., 2020) introduces cloze-style multiple-choice inference with one million samples. Cross-domain contributions include **ScholarlyRead** (Saikh et al., 2020), **ChemLLMBench** (Guo et al., 2023), and **ScholarChemQA**

(Chen et al., 2025).

Document-level QA Document-level question answering extends beyond single-paragraph contexts to require reasoning over multiple paragraphs or entire documents. Datasets such as **COVID-QA** (Möller et al., 2020) and **QASPER** (Dasigi et al., 2021) provide multi-paragraph contexts, while **QASA** (Lee et al., 2023) focuses on classifying reasoning types. Large-scale, synthesis-focused datasets include **ScholarQABench** (Asai et al., 2024), **SciQAG-24D** (Wan et al., 2024), **PeerQA** (Baumgärtner et al., 2025), and **LitSearch** (Ajith et al., 2024). For deeper reasoning over full documents, datasets like **LitQA** (Lála et al., 2023) and **SCIDQA** (Singh et al., 2024a) extend evaluation to full-text comprehension.

Scientific Reasoning Benchmarks[1] such as **SCIBENCH** (Wang et al., 2024a), **LLM4MatBench** (Ndayishimiye Rubungo et al., 2024), **SciEval** (Sun et al., 2024), **Xiezhi** (Gu et al., 2024), and **SciKnowEval** (Feng et al., 2024) assess symbolic, quantitative, and multi-level scientific reasoning, extending evaluation beyond text retrieval to scientific problem-solving and conceptual application, primarily through text-based problem statements.

Scientific QA datasets reflect a trajectory from paragraph-level comprehension to document- and multi-document-level reasoning. The evolution mirrors the rapid advancement of model capabilities. With the growing reliance on generative models, challenges regarding hallucination and evidence grounding have intensified. Future datasets must therefore emphasize rigorous verifiability and reasoning depth, ensuring that models demonstrate grounded scientific reasoning rather than plausible-sounding hallucinations.

4 Multimodal Understanding Datasets

Comprehending scientific literature demands the integration of tightly coupled multimodal evidence, ranging from textual descriptions to structured tables, figures (charts, diagrams and images), and formulas. This presents unique challenges compared to general-domain vision-language tasks, requiring models to possess deep domain knowledge and the capacity for exact quantitative reasoning. This section surveys datasets dedicated to multimodal capabilities, focusing on tasks where textual context alone is insufficient for complete understanding.

4.1 Table Understanding

Table understanding extends beyond structural parsing, encompassing semantic interpretation, cell-level linking, and complex reasoning. Datasets such as **S2abEL** (Lou et al., 2023), **Wiki-TabNER** (Koleva et al., 2024), and **AxCell** (Kardas et al., 2020) emphasize semantic enrichment by linking table cells to named entities, concepts, and experimental results mentioned in the text. Reasoning-oriented datasets further evaluate deeper comprehension. For example, **LEGOBench** (Singh et al., 2024b) connects table cells with methods, footnotes, and experimental outcomes, while **SCITAT** (Zhang et al., 2025) and **SciTabQA** (Ghosh et al., 2024) provide comprehensive coverage of reasoning across hybrid tabular and textual content.

The evolution of these datasets reflects a shift from isolated extraction to context-aware reasoning. Contemporary evaluations prioritize factual consistency and the traceability of evidence back to textual sources, ensuring models can accurately interpret experimental data. Future work should concentrate on advanced quantitative reasoning and symbolic logic, enabling models to perform calculations and derive statistical conclusions from complex tables.

4.2 Figure Understanding

Figure Captioning Datasets such as **ACL-FIG** (Karishma et al., 2023), **SciFIBench** (Roberts et al., 2024), and **SciCap+** (Yang et al., 2023) link visual structures to captions and in-text mentions, supporting figure-aware retrieval and multimodal QA. **ArXivCap** (Li et al., 2024b) is a figure-caption dataset comprising 6.4M images and 3.9M captions, sourced from 572K ArXiv papers spanning various scientific domains.

Chart Understanding Charts summarize experimental results with multiple data series, domain-specific units, and analytical conventions. For chart understanding, datasets can be grouped by their focus. Some, like **CharXiv** (Wang et al., 2024b) and **SciGraphQA** (Li and Tajbakhsh, 2023), connect charts to their textual context to support multimodal reasoning. In parallel, bench-

[1]Benchmarks are usually datasets designed for performance evaluation. We use the term here to follow the naming conventions found in the original papers or dataset names.

marks like **CHART-Info 2024** (Davila et al., 2024) provide a real-world evaluation standard. Sourced from PubMed Central, it defines seven hierarchical tasks from classification, data extraction to systematically assess comprehensive chart understanding. Evaluation metrics emphasize numerical accuracy, correct visual grounding, and textual-graphical coherence.

Diagram Understanding Diagrams (e.g., flowcharts, pipeline schematics, system architectures) encode procedural logic and symbolic relations that are intricately tied to the scientific paper context. **FlowLearn** (Pan et al., 2024) targets flowchart understanding with a scientific subset sourced from scientific literature and supports understanding-oriented tasks such as OCR-aware component interpretation and Visual Question Answering. Complementing this, **Plot2XML** (Cui et al., 2025) provides gold-standard mxGraph XML annotations for real scientific diagrams, enabling evaluation of whether models can translate diagram semantics into executable code. **SciDoc2Diagrammer-MAF** (Mondal et al., 2024) explores diagram generation grounded in document text, a task that implicitly tests a model's ability to synthesize structured content from the paper to produce faithful visual representations.

Image Understanding Datasets such as **MedICaT** (Subramanian et al., 2020) and **ROCOv2** (Rückert et al., 2024) facilitate fine-grained biomedical image-text understanding through caption- and concept-grounded supervision. Specifically, MedICaT prioritizes figure-text alignment and retrieval by incorporating inline textual references and subfigure-subcaption mappings. In contrast, ROCOv2 pairs radiology images with captions and specific medical concepts to enable tasks like caption prediction and classification. **Open-PMC-18M** (Baghbanzadeh et al., 2025) and **BIOMEDICA** (Lozano et al., 2025) expand on these foundations by significantly scaling up coverage within the scientific literature, thereby supporting more robust training and evaluation of scientific image understanding models.

The evolution of figure understanding datasets demonstrates a progression from simple visual-text alignment to deep, context-dependent reasoning. Across diverse modalitiesranging from data-rich charts and procedural diagrams to biomedical imagescontemporary benchmarks emphasize the critical role of grounding visual semantics in the accompanying scientific narrative. This shift ensures models can move beyond surface-level recognition to synthesize structure, procedural logic, and experimental evidence. Future work should concentrate on bridging the gap between visual perception and executable logic, enabling models to not only interpret but also generate and reason over complex scientific imagery with high fidelity.

4.3 Formula Understanding

Formulas in scientific literature encode complex mathematical and logical relationships that are challenging for machines. Understanding them requires grasping symbolic notation, structure, and computation. Datasets like **FormulaReasoning** (Li et al., 2025) provide annotated question-answer pairs for numerical reasoning, while **MATP-BENCH** (He et al., 2025) adds multimodal theorem-solving requiring visual-symbolic integration.

Effective formula understanding hinges on disambiguating symbols, contextual alignment, and enabling precise computation. Current datasets facilitate structural accuracy and stepwise reasoning, but future research should emphasize robust evaluation, cross-domain generalization, and incorporation of domain-specific constraints.

4.4 Multimodal Reasoning

Advancing document-level understanding necessitates a shift from isolated modality perception to structural cross-modal integration. Unlike datasets focusing on single figures or tables, **SciMMIR** (Wu et al., 2024) addresses the complex interplay between textual citations and multimodal content. It establishes a specialized benchmark for fine-grained alignment, requiring models to perform provenance-aware retrieval and verify the explicit semantic links between body text and diverse visual evidence. At the broadest level, benchmarks like **MMMU** (Yue et al., 2024) assess cross-disciplinary knowledge integration. Collectively, these efforts underscore persistent challenges in regional grounding, measurement fidelity, and semantic integration across heterogeneous scientific modalities.

Recent datasets like **MMSci** (Li et al., 2024c) and **SPIQA** (Pramanick et al., 2024) address paper-level multimodal reasoning, requiring models to aggregate distributed visual evidence across

sections. Pushing further, the **MathVerse** benchmark (Zhang et al., 2024c) focuses on geometric and mathematical reasoning in visual problem-solving.

Multimodal understanding datasets increasingly emphasize document-grounded integration across text, tables, figures, and formulas, moving beyond isolated extraction toward context-aware, evidence-traceable reasoning. They also push for precise quantitative and symbolic computation, along with mutual grounding across modalities, enabling more verifiable reasoning at both document and corpus scale.

5 Pre-training and Instruction Fine-tuning Datasets

The capabilities of scientific large language models (LLMs) is fundamentally shaped by the corpora used for pre-training and instruction fine-tuning. Pre-training enables models to acquire broad linguistic competence, domain knowledge, and reasoning priors from large-scale scientific text, whereas instruction fine-tuning guides model behavior toward generating outputs that are accurate, contextually grounded, and consistent with scientific conventions. This section surveys representative datasets supporting these two complementary stages in the development of scientific LLMs.

5.1 Foundational Pre-training Datasets

Foundational pre-training relies on large-scale scientific datasets that expose models to general linguistic structures, specialized terminology, and discipline-specific discourse patterns. Repositories such as **arXiv** (Cornell University, 2023) and **PubMed/PMC** (U.S. National Library of Medicine, 1996; National Center for Biotechnology Information, 2023) offer structured, peer-reviewed content spanning a wide range of scientific fields, forming reliable bases for technical vocabulary learning and domain-consistent language modeling. Extended resources such as **S2ORC** (Lo et al., 2020) further enrich this landscape by providing millions of full-text articles augmented with citation graphs, allowing models to capture intertextual dependencies and relational patterns that underpin scientific reasoning. More general multi-domain datasets, including **The Pile** (Gao et al., 2020), **RedPajama** (Weber et al., 2024), and **Dolma** (Soldaini et al., 2024), integrate scientific literature with web data and code, balancing domain specificity with linguistic diversity and robustness. Crisis-driven collections such as **CORD-19** (Wang and et al., 2020) demonstrate the importance of rapid dataset construction in response to urgent scientific challenges. Formula-oriented datasets like **Nemotron-CC-Math** (Mahabadi et al., 2025), which extract LaTeX expressions via layout-aware pipelines, further strengthen mathematical and symbolic reasoning capabilities. In the multimodal setting, **Multimodal ArXiv** (Li et al., 2024b) provides aligned document renderings and source text, enabling joint modeling of visual and textual scientific signals.

Pre-training datasets constitute the backbone of scientific LLMs by providing linguistic breadth and intertextual context. Recent trends emphasize richer structural annotations and heterogeneous representations, underscoring the importance of combining lexical coverage with inter-document relational knowledge for effective scientific pre-training. Key dataset attributes include disciplinary diversity, full-text availability, citation connectivity, and temporal span. Future datasets should pursue scale while incorporating structural and multimodal richness, capturing discourse organization, argumentation patterns, and cross-domain reasoning signals essential for robust scientific language modeling.

5.2 Instruction Fine-tuning Datasets

Instruction fine-tuning adapts pre-trained LLMs to produce outputs that conform to scientific norms, such as clarity, reproducibility, and evidence-based reasoning. Datasets including **SciInstruct** (Zhang et al., 2024a) and **SciRIFF** (Wadden et al., 2024) define structured tasks that emulate question answering, explanation generation, and hypothesis assessment, thereby aligning model behavior with domain-specific expectations. More specialized instruction corpora, such as **SMolInstruct** (Yu et al., 2024) and **HoneyBee** (Song et al., 2023), target molecular and materials science reasoning, reflecting a shift toward fine-grained, task-realistic instruction design. The incorporation of multilingual and socio-technical resources, exemplified by **CSL** (Li et al., 2022) and **IDL** (University of California, San Francisco, 2023), further improves robustness and applicability across diverse scientific settings. In addition, large-scale vision-language resources such as **BIOMEDICA** (Lozano et al.,

2025) provide millions of expert-curated biomedical imagetext pairs, supporting both instruction tuning and evaluation in multimodal scientific contexts. As mentioned in Section **??**, **IEPile** (Gui et al., 2024b) and **InstructIE** (Gui et al., 2024a) are specialized instruction fine-tuning datasets tailored for information extraction tasks.

Instruction fine-tuning datasets bridge the gap between generic language modeling and domain-aware scientific reasoning. Core design considerations include task realism, schema consistency, multilingual support, and adherence to scientific standards of explanation and evidence. Future directions should emphasize broader thematic coverage, more diverse instruction formats, and the integration of provenance and verification signals, enabling instruction corpora to more reliably guide LLMs toward accurate and context-sensitive scientific outputs.

6 Future Directions

Despite the rapid growth in scientific literature datasets, persistent challenges remain in annotation scalability, reasoning depth, cross-document inference, and the standardization of AI-ready corpora.

Scalable Annotation via Human-AI Synergy The complexity of scientific domains makes reliance on manual labeling unsustainable. Future frameworks should leverage LLMs to synthesize preliminary annotations at scale, employing experts primarily for verifying high-uncertainty instances and defining reasoning rubrics. Approaches like hybrid machine-human labeling (Wang et al., 2024a) and uncertainty-aware active learning pipelines (Liu et al., 2025; Wadden et al., 2024) offer a path to maximize efficiency without compromising fidelity. Furthermore, establishing community-wide standards and structured metadata (Lo et al., 2020; Zhang et al., 2024a) is essential to foster semantic consistency and cross-disciplinary reproducibility.

Authentic Multimodal Understanding To support deep, context-aware inference, the next generation of datasets should unify multimodal information into holistic frameworks. While recent works have begun to integrate heterogeneous inputs (Li et al., 2024c), future datasets should prioritize fine-grained semantic linkingtracing connections between textual claims and specific chart components or table cells (Wang et al., 2024b). Ultimately, datasets must simulate rigorous cross-modal synthesis, moving beyond retrieval to mandate the joint symbolic and visual reasoning (Pramanick et al., 2024) reflective of authentic scientific workflows.

Cross-document Scientific Reasoning Synthesizing evidence across dispersed literature is central to scientific discovery. Future datasets should expand beyond single-document tasks to require multi-hop inference over vast corpora, evaluating model ability to connect distant concepts (Lála et al., 2023). Crucially, models must learn to reconcile conflicting results across studies (Fernandes et al., 2024) and perform longitudinal analysis to track knowledge evolution. Leveraging temporal datasets (Liu et al., 2023b; Chen et al., 2025) will be key to enabling forward-looking reasoning and identifying emerging research gaps.

Standardized and Executable Ecosystems To resolve issues of reproducibility and interoperability, the field requires standardized, well-documented datasets. Future efforts should focus on consistent preprocessing protocols that systematically extract text, tables, figures and formulas (Pfitzmann et al., 2022), adaptable to diverse domains. Moreover, integrating rich metadata and traceable provenance (Lo et al., 2020; Hu et al., 2024) is vital for auditability. By establishing interoperable formats compatible with large-scale training (Smock et al., 2022; Singh et al., 2024b), future datasets will serve as the bedrock for robust, cross-domain AI systems.

7 Conclusions

Scientific literature understanding has become an imperative for coping with the unprecedented proliferation of scholarly publications. In this survey, we systematize the dataset landscape through a four-part taxonomy: structural parsing, text understanding, multimodal understanding and pre-training/instruction fine-tuning. This survey is intended to serve as a practical reference for dataset selection and a conceptual scaffold for future dataset construction, ultimately solidifying robust scientific literature understanding as a cornerstone for realizing the AI4S vision.

Limitations

This survey has several limitations. First, our scope is centered on datasets derived from scientific literature, and we do not comprehensively cover complementary AI4S resources such as experimental measurements, laboratory notebooks, simulations, molecular/protein databases, or instrument-generated multimodal signals, which are increasingly integrated into end-to-end scientific workflows. Second, the surveyed datasets are skewed toward English and high-resource disciplines, leaving multilingual settings and long-tail scientific domains underrepresented; consequently, conclusions about generalization may not fully transfer to low-resource languages or specialized subfields. Third, due to rapid dataset turnover and heterogeneous reporting standards, comparisons across datasets are sometimes impeded by inconsistent task definitions, annotation conventions, and evaluation protocols.

References

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. 2024. Litsearch: A retrieval benchmark for scientific literature search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15068–15083.

Dan Anitei, Joan Andreu Sánchez, José Miguel Benedí, and Ernesto Noya. 2023. The ibem dataset: A large printed scientific image dataset for indexing and searching mathematical expressions. *Pattern Recognition Letters*, 172:29–36.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, and 1 others. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*.

Negin Baghbanzadeh, Sajad Ashkezari, Elham Dolatabadi, and Arash Afkanpour. 2025. Openpmc-18m: A high-fidelity large scale medical dataset for multimodal representation learning. *Preprint*, arXiv:2506.02738.

Tim Baumgärtner, Ted Briscoe, and Iryna Gurevych. 2025. Peerqa: A scientific question answering dataset from peer reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 508–544.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. Tldr: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777.

Xiuying Chen, Tairan Wang, Taicheng Guo, Kehan Guo, Juexiao Zhou, Haoyang Li, Zirui Song, Xin Gao, and Xiangliang Zhang. 2025. Unveiling the power of language models in chemical research question answering. *Communications Chemistry*, 8(1):4.

Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. 2023. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15138–15147.

Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. 2019. Complicated table structure recognition (scitsr). *arXiv preprint arXiv:1908.04729*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Cornell University. 2023. arxiv. https://arxiv.org/. Online repository of electronic preprints.

Zhiqing Cui, Jiahao Yuan, Hanqing Wang, Yanshu Li, Chenxu Du, and Zhenglong Ding. 2025. Draw with thought: Unleashing multimodal reasoning for scientific diagram generation. *Preprint*, arXiv:2504.09479.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610.

Kenny Davila, Rupak Lazarus, Fei Xu, Nicole Rodríguez Alcántara, Srirangaraj Setlur, Venu Govindaraju, Ajoy Mondal, and CV Jawahar. 2024. Chartinfo 2024: A dataset for chart analysis and recognition. In *International Conference on Pattern Recognition*, pages 297–315. Springer.

Yuntian Deng, Anssi Kanervisto, Jonathan Ling, and Alexander M. Rush. 2016. Image-to-markup generation with coarse-to-fine attention. *arXiv preprint arXiv:1609.04938*.

Harsh Desai, Pratik Kayal, and Mayank Singh. 2021. Tablex: a benchmark dataset for structure and content information extraction from scientific tables. In *International Conference on Document Analysis and Recognition*, pages 554–569. Springer.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. 2024. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*.

Leandro Carísio Fernandes, Gustavo Bartz Guedes, Thiago Soares Laitz, Thales Sales Almeida, Rodrigo Nogueira, Roberto Lotufo, and Jayr Pereira. 2024. Surveysum: A dataset for summarizing multiple scientific articles into a survey section. In *Brazilian Conference on Intelligent Systems*, pages 431–444. Springer.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Liangcai Gao, Yilun Huang, Hervé Déjean, Jean-Luc Meunier, Qinqin Yan, Yu Fang, Florian Kleber, and Eva Lang. 2019. Icdar 2019 competition on table detection and recognition (ctdar). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1510–1515.

Andrea Gemelli, Simone Marinai, Lorenzo Pisaneschi, and Francesco Santoni. 2024. Datasets and annotations for layout analysis of scientific articles. *International Journal on Document Analysis and Recognition (IJDAR)*, 27(4):683–705.

Akash Ghosh, Venkata Sahith Bathini, Niloy Ganguly, Pawan Goyal, and Mayank Singh. 2024. How robust are the QA models for hybrid scientific tabular data? a study using customized dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8258–8264, Torino, Italia. ELRA and ICCL.

Felipe Grijalva, Carla Parra, Marco Gallardo, Erick Santos, Byron Acuña, Juan Carlos Rodríguez, and Julio Larco. 2022. Scibank: A large dataset of annotated scientific paper regions for document layout analysis.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, and 1 others. 2024. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18099–18107.

Honghao Gui, Shuofei Qiao, Jintian Zhang, Hongbin Ye, Mengshu Sun, Lei Liang, Jeff Z. Pan, Huajun Chen, and Ningyu Zhang. 2024a. Instructie: A bilingual instruction-based information extraction dataset. In *International Semantic Web Conference*.

Honghao Gui, Lin Yuan, Hongbin Ye, Ningyu Zhang, Mengshu Sun, Lei Liang, and Huajun Chen. 2024b. Iepile: Unearthing large-scale schema-based information extraction corpus. *arXiv preprint arXiv:2402.14710*.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What indeed can gpt models do in chemistry? a comprehensive benchmark on eight tasks. *Preprint*, arXiv:2305.18365.

Mark A Hanson, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington. 2024. The strain on scientific publishing. *Quantitative Science Studies*, 5(4):823–843.

Zhitao He, Zongwei Lyu, Dazhong Chen, Dadi Guo, and Yi R. Fung. 2025. Matp-bench: Can mllm be a good automated theorem prover for multimodal problems? *Preprint*, arXiv:2506.06034.

Kai Hu and 1 others. 2024. Mathematical formula detection in document images. *Pattern Recognition*. ArxivFormula dataset for formula entity detection and relation extraction on arXiv PDFs.

Huawei Ji, Cheng Deng, Bo Xue, Zhouyang Jin, Jiaxin Ding, Xiaoying Gan, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. 2025. Aceparse: A comprehensive dataset with diverse structured texts for academic literature parsing. In *ICASSP 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

KV Jobin, Ajoy Mondal, and CV Jawahar. 2019. Docfigure: A dataset for scientific document figure classification. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 1, pages 74–79. IEEE.

Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. 2020. Axcell: Automatic extraction of results from machine learning papers. *arXiv preprint arXiv:2004.14356*.

Zeba Karishma, Shaurya Rohatgi, Kavya Shrinivas Puranik, Jian Wu, and C Lee Giles. 2023. Acl-fig: A dataset for scientific figure classification. *arXiv preprint arXiv:2301.12293*.

JD Kim, T Ohta, Y Tateisi, and J Tsujii. 2003. Genia corpus–semantically annotated corpus for bio-textmining. *Bioinformatics (Oxford, England)*, 19:i180–i182.

Aneta Koleva, Martin Ringsquandl, Ahmed Hatem, Thomas Runkler, and Volker Tresp. 2024. Wikitabner: Advancing table interpretation through named entity recognition. *arXiv preprint arXiv:2403.04577*.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, and 1 others. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(Suppl 1):S2.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *CoRR*.

Esther Landhuis. 2016. Scientific literature: Information overload. *Nature*, 535(7612):457–458.

Peder Larsen and Markus Von Ins. 2010. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84(3):575–603.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. 2023. Qasa: advanced question answering on scientific articles. In *International Conference on Machine Learning*, pages 19036–19052. PMLR.

Hao Li, Peng Yue, Deodato Tapete, Francesca Cigna, Qiuju Wu, Longgang Xiang, and Binbin Lu. 2024a. Esdc: An open earth science data corpus to support geoscientific literature information extraction. *SCIENCE CHINA Earth Sciences*, 67(12).

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024b. Multimodal arxiv: A dataset for improving scientific comprehension of large vision–language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), Long Papers*, pages 14369–14387, Bangkok, Thailand.

Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. 2020a. Tablebank: Table benchmark for image-based table detection and recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1918–1925.

Minghao Li, Yi Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020b. Docbank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 949–960. International Committee on Computational Linguistics.

Shengzhi Li and Nima Tajbakhsh. 2023. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349*.

Xiao Li, Bolin Zhu, Kaiwen Shi, Sichen Liu, Yin Zhu, Yiwei Liu, and Gong Cheng. 2025. Formulareasoning: A dataset for formula-based numerical reasoning. *Preprint*, arXiv:2402.12692.

Yudong Li, Yuqing Zhang, Zhe Zhao, Linlin Shen, Weijie Liu, Weiquan Mao, and Hui Zhang. 2022. Csl: A large-scale chinese scientific literature dataset. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3917–3923.

Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, and 1 others. 2024c. Mmsci: A dataset for graduate-level multi-discipline multimodal scientific understanding. *arXiv preprint arXiv:2407.04903*.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023a. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10381–10399, Toronto, Canada. Association for Computational Linguistics.

Javin Liu, Aryan Vats, and Zihao He. 2025. Cs-papersum: A large-scale dataset of ai-generated summaries for scientific papers. *arXiv preprint arXiv:2502.20582*.

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2023b. Generating a structured summary of numerous academic papers: Dataset and method. *arXiv preprint arXiv:2302.04580*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983.

Yuze Lou, Bailey Kuehl, Erin Bransom, Sergey Feldman, Aakanksha Naik, and Doug Downey. 2023. S2abel: a dataset for entity linking from scientific tables. *arXiv preprint arXiv:2305.00366*.

Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Anita Rau, Austin Wolfgang Katzer, and 1 others. 2025. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19724–19735.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074.

Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. 2021. Chartocr: Data extraction from charts images via a deep hybrid framework. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1917–1925.

L Luo, PT Lai, CH Wei, CN Arighi, and Z Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282–bbac282.

Rabeeh Karimi Mahabadi, Sanjeev Satheesh, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Nemotron-cc-math: A 133 billion-token-scale high quality math pretraining dataset. *Preprint*, arXiv:2508.15096.

Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: a faceted summarization dataset for long scientific documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1080–1089.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. Covid-qa: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Ishani Mondal, Zongxia Li, Yufang Hou, Anandhavelu Natarajan, Aparna Garimella, and Jordan Boyd-Graber. 2024. Scidoc2diagrammer-MAF: Towards generation of scientific diagrams from documents guided by multi-aspect feedback refinement. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13342–13375.

National Center for Biotechnology Information. 2023. Pubmed central. `https://www.ncbi.nlm.nih.gov/pmc/`. Online archive of biomedical and life sciences journal literature.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Aime Ndayishimiye Rubungo, Qinghua Liang, Yuwei Qin, Yuxuan Wang, Xinyi Wang, Yichi Zhang, Jinyang Jiang, and Xiang Chen. 2024. Llm4matbench: Benchmarking large language models for materials property prediction. *arXiv preprint arXiv:2411.00177*.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, and 1 others. 2025. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24838–24848.

Huitong Pan, Qi Zhang, Cornelia Caragea, Eduard Dragut, and Longin Jan Latecki. 2024. Flowlearn: Evaluating large vision-language models on flowchart understanding. *Preprint*, arXiv:2407.05183.

Dimitris Pappas, Ion Androutsopoulos, and Harris Papageorgiou. 2018. Bioread: A new dataset for biomedical reading comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2771–2776.

Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. 2020. Biomrc: A dataset for biomedical machine reading comprehension. In *Proceedings of the 19th SIGBioMed workshop on biomedical language processing*, pages 140–149.

B Pfitzmann, C Auer, M Dolfi, A S Nassar, and P W Staar. 2022. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *KDD*, pages 3743–3751.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. *Advances in Neural Information Processing Systems*, 37:118807–118833.

Yujie Qian, Jiang Guo, Zhengkai Tu, Zhening Sun, Yousef Jradeh, Zhiping Cai, Nikolay Teslya, Ming Li, Yunlong Zhang, Fabian Hase, Philippe Schwaller, and Alán Aspuru-Guzik. 2023. Molscribe: Robust molecular structure recognition with image-to-graph generation. *Journal of Chemical Information and Modeling*, 63(7):1925–1934.

Kohulan Rajan, Henning Otto Brinkhaus, M. Isabel Agea, Achim Zielesny, and Christoph Steinbeck. 2023. Decimer.ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. *Nature Communications*, 14(1):5045.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. *Advances in Neural Information Processing Systems*, 37:18695–18728.

Johannes Rückert, Louise Bloch, Raphael Brüngel, and 1 others. 2024. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*.

Tanik Saikh, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Scholarlyread: A new dataset for scientific article reading comprehension. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5498–5504.

Shruti Singh, Nandan Sarkar, and Arman Cohan. 2024a. Scidqa: A deep reading comprehension dataset over scientific papers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20908–20923.

Shubham Singh, Yixuan Tang, Nicholas Monath, Andrew McCallum, Claire Cardie, Graham Neubig, Kyle Lo, Daniel S. Weld, Chandra Bhagavatula, and David Wadden. 2024b. Legobench: Scientific leaderboard generation benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14598–14613, Miami, USA.

Larry Smith, Lorraine K Tanabe, Rie Johnson Nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, and 1 others. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(Suppl 2):S2.

Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4634–4642, New Orleans, LA, USA.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 2 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Yu Song, Santiago Miret, Huan Zhang, and Bang Liu. 2023. Honeybee: Progressive instruction finetuning of large language models for materials science. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5724–5739.

Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Srinivasan Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. Medicat: A dataset of medical images, captions, and textual references. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2112–2120.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19053–19061.

University of California, San Francisco. 2023. Industry documents library. `https://www.industrydocuments.ucsf.edu/`. Online repository of documents related to industry.

U.S. National Library of Medicine. 1996. Pubmed. $https://pubmed.ncbi.nlm.nih.gov/$. Accessed: 2025-09-19.

David Wadden, Kejian Shi, Jacob Morrison, Aakanksha Naik, Shruti Singh, Nitzan Barzilay, Kyle Lo, Tom Hope, Luca Soldaini, Shannon Zejiang Shen, and 1 others. 2024. Sciriff: A resource to enhance language model instruction-following over scientific literature. *arXiv preprint arXiv:2406.07835*.

Yuwei Wan, Yixuan Liu, Aswathy Ajith, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu Kit, Tong Xie, and Ian Foster. 2024. Sciqag: A framework for auto-generated science question answering dataset with fine-grained evaluation. *arXiv preprint arXiv:2405.09939*.

Lucy Lu Wang and et al. 2020. Cord-19: The covid-19 open research dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8640–8646. Association for Computational Linguistics.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024a. Scibench: Evaluating college-level scientific problem-solving abilities of large language models.

In *International Conference on Machine Learning*, pages 50622–50649. PMLR.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024b. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

Maurice Weber, Dan Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, and 1 others. 2024. Redpajama: an open dataset for training large language models. *Advances in neural information processing systems*, 37:116462–116492.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Ranran Haoran Zhang, Bohao Yang, Wenhu Chen, and 1 others. 2024. Scimmir: Benchmarking scientific multimodal information retrieval. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12560–12574.

Yongjun Xu, Xin Liu, Xin Cao, Changping Huang, Enke Liu, Sen Qian, Xingchen Liu, Yanjun Wu, Fengliang Dong, Cheng-Wei Qiu, and 1 others. 2021. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 2(4).

Zihan Yang, Anran Wang, Shiyang Liu, and Kyle Lo. 2023. Scicap+: A knowledge-augmented dataset to study the faithfulness of scientific figure captioning. *arXiv preprint arXiv:2306.03491*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567. IEEE.

Dan Zhang, Ziniu Hu, Sining Zhoubian, Zhengxiao Du, Kaiyu Yang, Zihan Wang, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024a. Sciglm: Training scientific language models with self-reflective instruction annotation and tuning. *arXiv preprint arXiv:2401.07950*.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024b. Mm-llms: Recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12401–12430.

R. Zhang, D. Jiang, Y. Zhang, and 1 others. 2024c. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, ECCV, pages 169–186, Cham. Springer Nature Switzerland.

Weizhe Zhang and 1 others. 2024d. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*.

Xuanliang Zhang, Dingzirui Wang, Baoxin Wang, Longxu Dou, Xinyuan Lu, Keyan Xu, Dayong Wu, and Qingfu Zhu. 2025. Scitat: A question answering benchmark for scientific tables and text covering diverse reasoning types. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3859–3881, Vienna, Austria. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pages 564–580. Springer.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE.

A Appendix A: Dataset List

To facilitate future research and provide a clear overview of the data landscape, we have summarized the detailed statistics of existing datasets into four categorized tables. Specifically, Table 1 presents datasets focused on structural parsing, covering layout analysis and element recognition. Table 2 details resources for text understanding tasks, such as information extraction and summarization. Addressing more complex scenarios, Table 3 lists datasets for multimodal understanding. Finally, Table 4 provides specifications for datasets used in foundational pre-training and instruction fine-tuning.

Table 1: Datasets for Structural Parsing. **Abbreviations**: Multi-doc: Multi-document; Eval-only: Evaluation-only; Auto: Automatically Labeled; Human: Human-annotated.

Dataset	Year	Label	Domain	Size
Layout Analysis and End-to-end Parsing				
PubLayNet (Zhong et al., 2019)	2019	Auto	Biomedical	360k pages
DocBank (Li et al., 2020b)	2020	Auto	Computer Science	500k pages
SciBank (Grijalva et al., 2022)	2022	Human&Auto	Multi-domain	74,435 pages
DocLayNet (Pfitzmann et al., 2022)	2022	Human	Multi-domain	80k pages
OmniDocBench (Ouyang et al., 2025)	2024	Human	Multi-domain	981 pages
DocGenome (Zhang et al., 2024d)	2024	Auto	Multi-domain	6.8M pages
M6Doc	2022	Human	Multi-domain	237k instances
Nougat	2023	Auto	Multi-domain	8,494,841 pages
AceParse	2023	Auto	Multi-domain	500k parsed pairs
Table Recognition				
SciTSR (Chi et al., 2019)	2019	Auto	Multi-domain	15,000 examples
PubTabNet (Zhong et al., 2020)	2020	Auto	Multi-domain	568k tables
ICDAR-2019 (cTDaR) competition (Gao et al., 2019)	2019	Human	Multi-domain	1639 tables
PubTables-1M (Smock et al., 2022)	2021	Auto	Multi-domain	1M tables
TableBank (Li et al., 2020a)	2020	Auto	Multi-domain	417k tables
TabLeX (Desai et al., 2021)	2020	Auto	Multi-domain	4M images
Figure Recognition				
DocFigure (Jobin et al., 2019)	2019	Human&Auto	Multi-domain	33K figures
ChartOCR (Luo et al., 2021)	2021	Auto	Computer Science	386,966 chart images
DePlot (Liu et al., 2023a)	2023	Auto	Multi-domain	516k parsed pairs
Formula Recognition				
IM2LATEX (Deng et al., 2016)	2017	Auto	Multi-domain	103,556 equations
ArxivFormula (Hu et al., 2024)	2023	Auto	Multi-domain	600k pages
IBEM (Anitei et al., 2023)	2022	Human	Multi-domain	166k expressions
DECIMER.ai (Rajan et al., 2023)	2023	Auto	Chemistry	450M depictions
MolScribe (Qian et al., 2023)	2023	Human	Scientific	331 images

Table 2: Datasets for Text Understanding. **Abbreviations**: Multi-doc: Multi-document; Eval-only: Evaluation-only; NER: Named Entity Recognition; RE: Relation Extraction; SDS: single-document summarization; MDS: Multi-Document Summarization; Para-QA: Paragraph-level Question Answering; Doc-QA: Document-level Question Answering; Auto: Automatically Labeled; Human: Human-annotated.

Dataset	Year	Task	Multi-doc	Eval-only	Label	Domain	Size
Information Extraction							
GENIA (Kim et al., 2003)	2003	NER	No	No	Human	Biomedical	2,000 abstracts
BC2GM (Smith et al., 2008)	2008	NER	No	No	Human	Biomedical	20k sentences
NCBI Disease (Doğan et al., 2014)	2014	NER	No	No	Human	Biomedical	800 abstracts
CHEMDNER (Krallinger et al., 2015)	2015	NER	No	No	Human	Biomedical	10k abstracts
ESDC (Li et al., 2024a)	2023	IE	-	No	Human	Earth Science	500 abstracts
BC5CDR (Li et al., 2016)	2016	RE	No	No	Human	Biomedical	1.5k abstracts
BioRED (Luo et al., 2022)	2022	RE	No	No	Human	Biomedical	600 abstracts
IEPile (Gui et al., 2024b)	2024	NER	No	No	Human&Auto	Multi-domain	2M instances
InstructIE (Gui et al., 2024a)	2024	IE	No	No	Human&Auto	Multi-domain	364k instances
Summarization							
WikiHow (Koupaee and Wang, 2018)	2018	SDS	No	No	Human	Multi-domain	230k articles
PubMed/Arxiv Summarization (Cohan et al., 2018)	2018	SDS	No	No	Auto	Biomedical	348k papers
SCITLDR (Cachola et al., 2020)	2020	SDS	No	No	Human	Multi-domain	5.4k full-text papers
FacetSum (Meng et al., 2021)	2021	SDS	No	No	Human	Multi-domain	60k full-text papers
CS-PaperSum (Liu et al., 2025)	2025	SDS	No	No	Auto	Computer Science	91.9k full-text papers
Multi-XScience (Lu et al., 2020)	2020	MDS	Yes	No	Auto	Multi-domain	30k instances
MS^2 (DeYoung et al., 2021)	2021	MDS	Yes	No	Human	Biomedical	20k reviews
BigSurvey (Liu et al., 2023b)	2022	MDS	Yes	No	Auto	Multi-domain	7k full-text surveys
SurveySum (Fernandes et al., 2024)	2024	MDS	Yes	No	Human	Multi-domain	79 sections
Question Answering							
BioRead (Pappas et al., 2018)	2018	Para-QA	No	No	Human&Auto	Biomedical	10k QA pairs
PubMedQA (Jin et al., 2019)	2019	Para-QA	No	No	Human&Auto	Biomedical	273k QA pairs
BioMRC (Pappas et al., 2020)	2020	Para-QA	No	No	Auto	Biomedical	1M QA pairs
ScholarlyRead (Saikh et al., 2020)	2020	Para-QA	No	No	Human	Multi-domain	10k QA pairs
ChemLLMBench (Guo et al., 2023)	2023	Para-QA	No	Yes	Human	Chemistry	8 tasks
ScholarChemQA (Chen et al., 2025)	2025	Para-QA	No	No	Human&Auto	Chemistry	40k QA pairs
COVID-QA (Möller et al., 2020)	2020	Doc-QA	Yes	No	Human	COVID-19	2,019 QA pairs
QASPER (Dasigi et al., 2021)	2021	Doc-QA	Yes	No	Human	NLP	5k QA pairs
QASA (Lee et al., 2023)	2023	Doc-QA	Yes	No	Human	Multi-domain	1.8k QA pairs
ScholarQABench (Asai et al., 2024)	2024	Doc-QA	Yes	Yes	Human	Multi-domain	2.9k QA pairs
SciQAG-24D (Wan et al., 2024)	2024	Doc-QA	Yes	No	Auto	Multi-domain	188k QA pairs
PeerQA (Baumgärtner et al., 2025)	2025	Doc-QA	Yes	No	Human	Multi-domain	579 QA pairs
LitSearch (Ajith et al., 2024)	2024	Doc-QA&Retrieval	Yes	No	Human	Multi-domain	597 queries
LitQA (Lála et al., 2023)	2023	Doc-QA	Yes	No	Human	Biology	50 QA pairs
SCIDQA (Singh et al., 2024a)	2024	Doc-QA	Yes	No	Human	Multi-domain	2.9k QA pairs
SCIBENCH (Wang et al., 2024a)	-	Reasoning	No	Yes	Human	Multi-domain	695 samples
LLM4Mat-Bench (Ndayishimiye Rubungo et al., 2024)	2024	Reasoning	No	Yes	Auto	Materials Science	1.97M samples
SciEval (Sun et al., 2024)	2024	Reasoning	No	Yes	Human	Multi-domain	18k QA pairs
Xiezhi (Gu et al., 2024)	2024	Reasoning	No	Yes	Human	Multi-domain	249k QA pairs
SciKnowEval (Feng et al., 2024)	2024	Reasoning	No	Yes	Human	Multi-domain	70k QA pairs

Table 3: Datasets for Multimodal Understanding. **Abbreviations**: Multi-doc: Multi-document; Eval-only: Evaluation-only; Auto: Automatically Labeled; Human: Human-annotated.

Dataset	Year	Multi-doc	Eval-only	Label	Domain	Size
Table Understanding						
S2abEL (Lou et al., 2023)	2023	No	No	Human&Auto	Computer Science	732 tables
Wiki-TabNER (Koleva et al., 2024)	2024	No	No	Auto	Multi-domain	61,235 tables
AxCell (Kardas et al., 2020)	2020	No	No	Human	Multi-domain	over 100,000 papers
LEGOBench (Singh et al., 2024b)	2024	Yes	Yes	Human	Computer Science	3,666 datasets
SCITAT (Zhang et al., 2025)	2025	No	No	Human	Computer Science	953 questions
SciTabQA (Ghosh et al., 2024)	2024	No	No	Human	Computer Science	198 tables and 822 QA pairs
Figure Understanding						
ACL-FIG (Karishma et al., 2023)	2023	No	No	Human	Comp. Sci. (NLP)	112,052 figures
SciFIBench (Roberts et al., 2024)	2024	No	Yes	Human	Scientific	2,000 questions
SciCap+ (Yang et al., 2023)	2023	No	No	Human	Scientific	414,809 figures
ArXivCap (Li et al., 2024b)	2024	No	No	Auto	Scientific	6.4M images & 3.9M captions
CharXiv (Wang et al., 2024b)	2024	No	No	Human	Scientific	2,323 charts
SciGraphQA (Li and Tajbakhsh, 2023)	2023	No	No	Human	Scientific	295K samples
CHART-Info 2024 (Davila et al., 2024)	2024	No	Yes	Human	Scientific	3.2K charts
FlowLearn (Pan et al., 2024)	2024	No	No	Human&Auto	Multi-domain	13,858 diagrams
Plot2XML (Cui et al., 2025)	2025	No	Yes	Human	Computer Science	247 diagrams
SciDoc2Diagrammer-MAF (Mondal et al., 2024)	2024	Yes	No	Human&Auto	Computer Science	1080 diagrams
MedICaT (Subramanian et al., 2020)	2020	No	No	Human&Auto	Biomedical	217,060 images
ROCOv2 (Rückert et al., 2024)	2024	No	No	Human&Auto	Biomedical	79,789 images
Open-PMC-18M (Baghbanzadeh et al., 2025)	2025	No	No	Human&Auto	Biomedical	18M image-text pairs
BIOMEDICA (Lozano et al., 2025)	2025	No	No	Human&Auto	Biomedical	30,711,542 Figure referenecs
Formula Understanding						
FormulaReasoning (Li et al., 2025)	2025	No	No	Human	Scientific	5,324 questions
MATP-BENCH (He et al., 2025)	2025	No	Yes	Human	Math	1056 multimodal theorems
Multimodal Reasoning						
SciMMIR (Wu et al., 2024)	2024	No	No	Human	Scientific	530K image-text pairs
MMMU (Yue et al., 2024)	2023	No	Yes	Human	Multi-domain	11.5K questions
MMSci (Li et al., 2024c)	2024	No	No	Human	Scientific	131,393 articles
SPIQA (Pramanick et al., 2024)	2024	Yes	No	Human	Scientific	270K questions
MathVerse (Zhang et al., 2024c)	2024	No	Yes	Human	Math/Geometry	15K test samples

Table 4: Datasets for Pre-training and Instruction Fine-tuning. **Abbreviations**: Auto: Automatically Labeled; Human: Human-annotated.

Dataset	Year	Label	Domain	Size
Foundational Pre-training Datasets				
arXiv (Cornell University, 2023)	1991	Auto	Multi-domain	2.6M full-text papers
PubMed (U.S. National Library of Medicine, 1996)	1966	Auto	Biomedical	35M paper metadata
PubMed Central (PMC) (National Center for Biotechnology Information, 2023)	2000	Auto	Biomedical	10M full-text papers
S2ORC (Lo et al., 2020)	2020	Auto	Multi-domain	81M full-text papers
The Pile (Gao et al., 2020)	2020	Auto	Multi-domain	800GB raw text
RedPajama (Weber et al., 2024)	2024	Auto	General	30T Tokens
Dolma (Soldaini et al., 2024)	2024	Auto	General	3T tokens
CORD-19 (Wang and et al., 2020)	2020	Auto	COVID-19	1M full-text papers
Nemotron-CC-Math (Mahabadi et al., 2025)	2025	Human&Auto	Math	133B Tokens
Multimodal ArXiv (Li et al., 2024b)	2024	Human&Auto	Multi-domain	3M instances
Instruction Fine-tuning Datasets				
SciInstruct (Zhang et al., 2024a)	2024	Human	Multi-domain	100k examples
SciRIFF (Wadden et al., 2024)	2024	Human	Multi-domain	3.1k abstracts
SMolInstruct (Yu et al., 2024)	2024	Auto	Molecular	340k instructions
HoneyBee (Song et al., 2023)	2023	Human&Auto	Materials Science	52k instructions
CSL (Li et al., 2022)	2022	Auto	Multi-domain	396k paper metadata
IDL (University of California, San Francisco, 2023)	2023	Human	Industry	15M documents
BIOMEDICA (Lozano et al., 2025)	2025	Human&Auto	Biomedical	30,711,542 Figure referenecs
IEPile (Gui et al., 2024b)	2024	Human&Auto	Multi-domain	33 collective datasets
InstructIE(Gui et al., 2024a)	2024	Human&Auto	Multi-domain	364,076 instances