Scientific Table Generation: Structured Information Extraction from Scientific Literature

Necva Bölücü, Huichen Yang*, Roelien C. Timmer, Stephen Wan,
Brian Jin, Changhyun Lee, Andreas Duenser, Jessica Irons
CSIRO, Australia
{necva.bolucu, huichen.yang, roelien.timmer, stephen.wan,
brian.jin, aiden.lee, andreas.duenser, jessica.irons}@csiro.au

Abstract

We introduce *Scientific Table Generation (STG)*, the task of extracting information from scientific documents to populate a table. We argue that STG can serve as a unifying task for structured knowledge extraction from scientific literature. To support this goal, we present the first multi-disciplinary benchmark for STG with predefined schemas and multi-record extraction, covering six scientific domains and comprising 20 datasets unified under our proposed formulation. We detail the benchmark's construction, including dataset composition, and introduce an *STG evaluation protocol*. We analyse the results of zero-shot large language model (LLM) and vision-language model (VLM) pipelines with typical document processing AI methods for the STG task. In doing so, we identify key challenges of the STG task and propose a set of open challenges, including multimodality, table and figure understanding, normalisation of data, and treatment of reported quantities. All code and data accompanying this work are publicly available at[1]: https://github.com/bolucunecva/table_generation

1 Introduction

The exponential growth of scientific literature has made it increasingly challenging for researchers to extract and organise key findings efficiently (Cai et al., 2024; Reddy and Shojaee, 2025). Manual methods, once manageable, are now overwhelmed by the sheer volume of publications (Borah et al., 2017; Cai et al., 2024). This has created a niche for automated tools (e.g., Elicit[2] and SciLire (Bölücü et al., 2026)), that can extract and structure information from scientific texts.

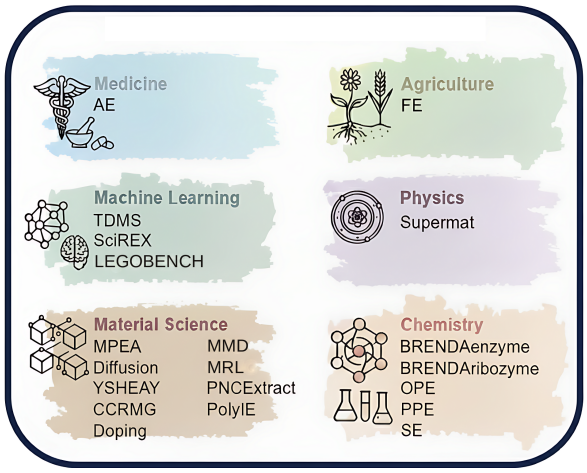


Figure 1: Benchmark covers six scientific domains: *Medicine*, *Agriculture*, *Machine Learning*, *Physics*, *Material Science*, and *Chemistry*.

Information extraction (IE) provides a foundational capability for this transformation by converting knowledge from unstructured sources into structured data (e.g., scientific publications, experimental notes). Such derived data can accelerate knowledge discovery (Horawalavithana et al., 2022), support predictive modelling (Li et al., 2022), and help in the development of AI-powered scientific assistants that can help researchers in summarising and reasoning about large scientific corpora (Lu et al., 2024; Su et al., 2024; Kulkarni et al., 2025).

While many recognise the need for IE solutions for scientific literature, there has not been a strong consensus in the problem framing. This has led to the exploration of this topic under various names within the Natural Language Processing (NLP) community, causing fragmentation and inconsistency across studies. Different terms have been used, such as joint named entity recognition and relation extraction (Townsend et al., 2021; Huguet Cabot and Navigli, 2021), n-ary relation

* This work was done when the author was affiliated with CSIRO.

[1]This repository will save researchers significant time. Instead of manually searching for 20 different datasets, researchers can use our repository to download all datasets directly, as a download link is provided for each paper. In addition, researchers can use our evaluation framework, which streamlines the process and improves overall efficiency.

[2]https://elicit.com

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 224–246
November 10, 2026 ©2026 Association for Computational Linguistics

extraction (Jia et al., 2019; Khalighinejad et al., 2024), Unified Text-to-Schema (Song et al., 2023), structured IE (Dunn et al., 2022), structured information inference (Xie et al., 2023b), ad-hoc schema-based IE (Ghosh et al., 2024), and instruct-and-extract (Jiao et al., 2023). These studies introduce separate datasets, differing in format and evaluation strategies, making it difficult to compare approaches or build systems applicable to all tasks. Moreover, most focus on specialised scientific domains, e.g., materials science (Ghosh et al., 2024; Cheung et al., 2024) and enzymes (Jiang et al., 2025).

We note that the existing common goal in these works remains the same: to extract structured data aligned with a predefined schema, a format into which the extracted information is organised. This is considered a comprehensive-level task (identifying, extracting and understanding the core content of documents) (Krathwohl, 2002) in scientific knowledge benchmarks (Feng et al., 2024; Cai et al., 2025).

In this work, we introduce the term **Scientific Table Generation** (STG), which formalises the task as extracting key values from unstructured scientific documents and organising them into a predefined tabular schema. Noting that scientists often use tables as a data format, we adopt a table-centric approach, as it offers a clear, consistent, and interpretable representation of scientific findings, enabling easier integration and downstream applications (Jiang et al., 2025; Wei et al., 2025).

To support research for this unified task, we curate the first *multi-disciplinary* benchmark (see Figure 1), standardising datasets introduced in prior work into a common table-centric format. For our selection criteria, we focused on (1) responsible AI datasets compliant with FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Wilkinson et al., 2016); (2) document-level extraction of records (records filled with information dispersed throughout a document); and (3) records with more than two fields.

Our assembled benchmark spans 6 scientific domains[3] and includes 20 datasets, reflecting the broad applicability and diversity across scientific domains. In addition, we document an evaluation framework for STG to help evaluate systems across different scientific domains. In this paper, we dive into the nature of the STG task, identifying specific challenges, e.g., inconsistencies in value formatting (e.g., units of measurement) and entity disambiguation. These insights lead us to propose STG open research topics.

The overarching goal of this paper is to establish a robust benchmark for STG, enabling consistent and rigorous evaluation of methodologies developed for this task. Correspondingly, our paper focuses on three primary contributions. First, we present the unified task formalism of STG (§ 3) and describe the benchmark (§ 4), whereby each dataset is reformatted to align with a common STG schema and includes: (i) *links* to documents in PDF format (e.g., DOI, PubMed, or ArXiv IDs); and (ii) ground truth tables aligned with the schema. Second, to facilitate comparisons, we also introduce an evaluation protocol (§ 5) that is customised for the STG task, such as mechanisms to handle challenges in scientific IE, such as varying units, format variations, and multiple records per document. Finally, our third contribution is a characterisation of the technical challenges of the STG task, as revealed by applying baseline LLM and VLM pipelines to the benchmark data. We analyse the inference results to gauge the relative differences across scientific domains and insights about the complexities of the task.

2 Related Work

Scientific IE STG is part of a broader set of IE tasks from scientific documents, such as extracting keyphrases and relations from publications (Augenstein et al., 2017), measuring science evolution via citation and entity extraction (Jurgens et al., 2018), classifying scientific discourse sentiments (Teufel et al., 2009), and automatically constructing scientific knowledge graphs (Ammar et al., 2018). These examples represent a small sample, with STG contributing by focusing on structured extraction and reconstruction of scientific tables reporting experimental results. ArxivDIGESTables (Newman et al., 2024) studies scientific table generation for literature-review synthesis, where both the schema and the values are generated with one row per paper

In our work, we focus on table-centric document IE, a more general problem that subsumes n-ary relation extraction. To focus on the more challenging general problem, we move on from the well-explored space of binary relation extraction

[3]Material Science appears as the dominant class, but is an interdisciplinary field spanning physics, chemistry, biology, and engineering.

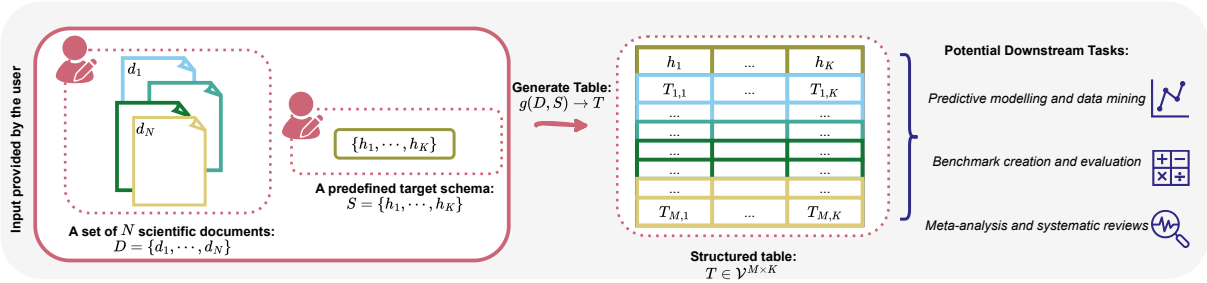


Figure 2: Visualisation of the Scientific Table Generation (STG) Task. First, the user provides N scientific documents $D = \{d_1, d_2, \cdots, d_N\}$ and a predefined target schema $\{h_1, \cdots, h_K\}$. Second, a structured table is generated: $g(D, S) \rightarrow T \in \mathcal{V}^{M \times K}$. The generated tables can be used for many downstream tasks, such as *predictive modelling and data mining*, *benchmark creation and evaluation*, and *meta-analysis and systematic reviews*.

and entity extraction, and specialised n-ary relation extraction instances that have been well explored in prior NLP work. That is, in favour of multi-field records (the rows in the table), we excluded binary relation datasets (e.g. SciERC (Luan et al., 2018) and DocRED (Yao et al., 2019)), abstract-level datasets (e.g. SemEval-2018 (Gábor et al., 2018), BioRED (Luo et al., 2022), and DWIE (Zaporojets et al., 2021)). We also exclude datasets targeting unrelated tasks such as classification or regression (e.g., Darwin (Xie et al., 2023a)).

Scientific Knowledge Benchmarks The presented benchmark is not the first benchmark proposed for evaluating scientific knowledge tasks. Other benchmarks include SciAssess (Cai et al., 2025), SciKnowEval (Feng et al., 2024), SciFact (Wadden et al., 2020), and SciBench (Wang et al., 2024a). SciAssess evaluates LLM performance on scientific literature analysis tasks, subdivided into *memorisation*, *comprehension*, and *analysis&reasoning*. SciKnowEval is similar, targeting scientific LLM tasks across difficulty levels: *studying extensively*, *inquiring earnestly*, *thinking profoundly*, *discerning clearly*, and *practising assiduously* (Feng et al., 2024). SciBench evaluates reasoning capabilities for complex scientific problems (Wang et al., 2024a), while SciFact focuses on verifying scientific claims based on abstracts (Wadden et al., 2020). In contrast, the *STG* benchmark uniquely targets the extraction of tables, which is not addressed by existing benchmarks.

3 Scientific Table Generation

Scientific Table Generation (STG) is the task of automatically generating a structured table of reported data from unstructured scientific publications conforming to a predefined schema, a structure that defines the format into which the extracted information is organised (Borg et al., 2020; Ghosh et al., 2024). The goal is to extract relevant information across different sections of a scientific document, representing these in tabular form, where each row is a scientific record, and each column corresponds to a schema attribute.

We formally describe the STG task as follows. Let $D = \{d_1, d_2, \cdots, d_N\}$ denote a collection of N scientific documents and $S = \{h_1, h_2, \cdots, h_K\}$ be a predefined target schema, where each h_j represents an attribute (or column) in the schema with a total of K columns. The STG task is the IE transformation, $g(D, S) \rightarrow T$, where T is a table of data of the form $T \in \mathcal{V}^{M \times K}$, where $M = \sum_{i=1}^{N} R_i$, where R_i is the total number of scientific records (i.e., rows) per document $d_i \in D$ that contributes to T.

Furthermore, each cell $T_{m,j}$ for row m (i.e. a single record) and column j, contains a string value, $V \in \mathcal{V}$. This value is a string that can be interpreted as a qualitative text answer, a reported quantity, or a category. Figure 2 shows a visualisation of the STG task, illustrating how information from scientific documents is mapped into a structured table consistent with the target schema. This formulation highlights two core challenges: (i) detecting and aggregating relevant evidence across potentially long and complex documents, and (ii) organising the extracted information into structured records that conform to a fixed schema.

4 Benchmark Construction

We introduce the benchmark, a suite of repurposed datasets initially proposed for varying related NLP tasks (e.g., IE, Q&A, . . .), covering 6 scientific domains: *Medicine*, *Agriculture*, *Machine Learning*, *Physics*, *Material Science*, and *Chemistry*.

Domain	Dataset	# Documents	Avg. Pages	# Records	Records / Document	Schema Size	Reference
Machine Learning	TDMS	332	10.52	904	2.72	4	Hou et al. (2019)
	SciREX	372	11.83	1,897	5.34	5	Jain et al. (2020)
	LEGOBENCH	6,995	14.63	145,419	20.79	5	Singh et al. (2024)
Material Science	MPEA	264	8.70	1,544	5.85	17	Borg et al. (2020)
	Diffusion	93	14.02	3,533	37.99	18	Zhang et al. (2010)
	YSHEAY	219	15.65	837	3.82	3	Polak and Morgan (2024)
	CCRMG	24	11.96	43	1.79	3	Polak and Morgan (2024)
	Doping	66	6.66	544	8.24	3	Dunn et al. (2022)
	MMD	3	6.00	140	46.67	16	Xie et al. (2023a)
	MRL	100	5.81	993	9.93	19	Xie et al. (2023a)
	PNCExtract	155	8.92	838	5.41	6	Khalighinejad et al. (2024)
	PolyIE	76	8.58	2,337	30.75	3	Cheung et al. (2024)
Chemistry	BRENDA-enzyme	155	12.45	4,210	27.16	13	Jiang et al. (2025)
	BRENDA-ribozyme	163	11.05	1,756	10.77	17	Jiang et al. (2025)
	OPE	104	7.55	255	2.45	7	Cai et al. (2025)
	PPE	109	6.97	265	2.43	6	Cai et al. (2025)
	SE	96	8.01	2,363	24.61	3	Cai et al. (2025)
Medicine	AE	40	4.30	406	10.15	3	Cai et al. (2025)
Agriculture	FE	326	11.65	326	1.00	15	Bölücü et al. (2025)
Physics	SuperMat	142	8.55	1,301	9.16	4	Foppiano et al. (2021)

Table 1: Statistics of 20 datasets across six domains. The schema of each dataset is given in the Appendix A in Table 3.

Dataset Selection Criteria This benchmark is built on datasets that meet several criteria: (i) document-level datasets (excluding sentence- or abstract-level IE data sets) and full PDFs can be obtained; (ii) adherence to the FAIR principles (Wilkinson et al., 2016) to support reproducibility and community use; and (iii) have n-ary relations that reflect the complexity of scientific knowledge.[4]

These datasets offer a blend of coarse- and fine-grained information, originating from either publication in PDF format or a structured document representation (e.g., XML) of a publication. For datasets that have only released structured representations (e.g., SciREX (Jain et al., 2020), MRL (Xie et al., 2023a)), we collect the corresponding PDFs via automatic scripts.

Design Choice For STG, we stipulate that the task should start with the PDF version of a publication. That is, a key design choice in building this benchmark is that we *do not force* a conversion of the PDFs into plain text, as this presumes some level of preprocessing that may become outdated in time. For example, as multimodal approaches become more advanced, we may *eventually* see a shift away from PDF parsing as a preprocessing step. To facilitate this, we distribute links to the PDF publication (e.g., DOI, ArXiv IDs). This decision has benefits: (i) the input can include non-textual elements (e.g., tables and figures) that are known to contain key information relevant to information reuse; and (ii) enables researchers to start with the original PDF, thereby encouraging transparency of approaches and reproducibility.[5]

Overview and Scope The final benchmark includes 20 datasets that meet our selection criteria. They cover a wide range of scientific concepts, for example, leaderboard construction (e.g., TDMS (Hou et al., 2019), LegoBench (Singh et al., 2024)), enzyme kinetics (e.g., BRENDA (Jiang et al., 2025)), and polymer collar cells (e.g., PolyIE (Cheung et al., 2024)), to name but a few research topics. The collection includes a variety of data types, covering both qualitative and quantitative information at various levels of detail. Table 1 provides an overview of the datasets included in the benchmark. Additional details on the schema structure, document sources, and domain can be found in Appendix A.

5 An Evaluation Protocol for STG

Evaluating IE from scientific documents presents unique challenges. Ideally, one would evaluate with human experts routinely (Dunn et al., 2022; Asai et al., 2024; Wang et al., 2024b); however, this is time-consuming and expensive. We note that scientists often judge the correctness of records as a

[4]Here we use the term n-ary relations to refer to a set of arguments, some of which might be subsumed by a common predicate while other arguments are merely associated as part of a discipline-specific IE task.

[5]Note that, due to copyright reasons, we do not distribute the PDFs as part of this dataset. Rather, we provide links and code for others to download the PDFs themselves.

Domain	Dataset	GPT-OSS				o1-mini				LLaMA 3.3				Nanonets (vision)			
		ChrF	Precision	Recall	F_1	ChrF	Precision	Recall	F_1	ChrF	Precision	Recall	F_1	ChrF	Precision	Recall	F_1
Machine Learning	TDMS	10.91	0.74	16.28	1.41	12.19	2.25	21.03	**4.07**	11.46	1.99	18.38	3.59	9.01	0.36	8.49	0.69
	SciREX	6.45	0.60	10.30	1.14	6.75	1.09	8.44	1.93	6.95	1.48	8.93	**2.54**	3.01	0.01	0.28	0.03
	LEGOBENCH	6.58	1.41	8.76	2.43	12.10	2.70	14.50	**4.55**	11.10	2.42	13.18	4.10	8.52	1.81	12.62	3.17
Material Science	MPEA	1.69	7.05	41.36	12.04	2.78	18.47	42.03	**25.66**	2.00	10.20	37.83	16.07	1.00	0.03	0.86	0.06
	Diffusion	0.74	1.12	19.06	2.12	0.94	22.67	13.20	**16.68**	0.88	10.52	10.16	10.34	0.44	1.06	10.46	1.92
	YSHEAY	13.19	0.57	23.58	1.11	16.31	2.86	44.29	**5.37**	14.25	1.82	34.37	3.47	3.99	1.40	2.51	1.80
	CCRMG	10.56	0.63	19.42	1.22	11.85	0.85	20.16	**1.63**	10.49	0.70	18.56	1.35	5.14	0.32	5.12	0.60
	Doping	10.19	4.83	20.15	7.79	14.96	23.30	19.16	**21.03**	12.52	13.38	20.83	16.29	2.17	1.21	2.77	1.69
	MMD	0.89	1.85	1.74	1.79	0.89	4.83	1.07	1.75	1.09	8.77	3.57	**5.08**	1.07	5.00	0.36	0.67
	MRL	0.57	1.28	3.00	1.79	0.73	2.86	1.25	1.74	0.72	2.44	2.82	**2.62**	0.71	0.38	2.22	0.65
	PNCExtract	8.74	6.52	44.15	11.37	8.64	23.96	39.14	**29.72**	9.40	11.73	27.96	16.53	5.22	0.42	13.17	0.81
	PolyIE	14.66	3.18	15.12	5.26	18.53	15.78	32.83	**21.31**	19.02	11.74	27.00	16.36	8.45	0.58	6.08	1.06
Chemistry	BRENDA-enzyme	2.39	11.01	27.64	15.75	2.55	24.30	26.67	**25.43**	2.53	11.07	30.72	16.28	1.21	0.75	8.22	1.38
	BRENDA-ribozyme	1.35	1.89	7.81	3.04	1.27	6.09	8.02	**6.92**	1.37	4.72	11.39	6.68	0.71	0.04	1.99	0.08
	OPE	3.90	2.06	34.40	3.88	4.75	7.08	38.60	**11.96**	4.76	4.66	37.70	8.29	2.99	0.07	10.81	0.14
	PPE	9.61	4.29	54.03	7.94	10.41	13.01	59.62	**21.37**	10.75	8.86	61.32	15.48	2.90	0.11	1.64	0.21
	SE	8.04	12.38	31.18	17.72	11.22	31.51	36.93	**34.01**	9.74	18.40	30.41	22.93	3.52	1.02	12.43	1.88
Medicine	AE	6.51	4.80	7.85	5.96	5.35	7.76	4.30	5.54	6.49	7.34	6.84	**7.08**	8.02	0.88	24.30	1.70
Agriculture	FE	2.22	0.85	36.58	1.67	2.52	1.89	28.45	**3.54**	2.50	1.49	28.20	2.83	1.40	1.63	5.46	2.52
Physics	SuperMat	6.53	11.91	26.84	16.50	10.05	33.28	29.72	**31.40**	9.04	12.49	19.23	15.15	2.94	1.70	8.46	2.83
	Avg.	6.29	3.95	22.46	6.10	7.74	12.33	24.47	**13.78**	7.35	7.31	22.47	9.65	3.62	0.94	6.91	1.19

Table 2: Performance comparison of models. All scores are reported on a 0–100 scale, with the best F_1 score for each dataset highlighted in **boldfaced**.

whole, rather than a component-based evaluation that focuses on cell contents in isolation. Existing evaluation procedures, moreover, generally do not account for correctness at the record or tuple level (Xie et al., 2023a; Ghosh et al., 2024; Cheung et al., 2024; Feng et al., 2024; Jiang et al., 2025). A notable exception is PNCExtract (Khalighinejad et al., 2024), which aligns predicted and reference records via maximum-weight bipartite matching before scoring. Our protocol builds on this idea and generalises record-level evaluation to the heterogeneous schemas of all datasets in the benchmark. We propose an evaluation protocol that assesses output with a record-level component.

We align predicted and reference records using the Hungarian Maximum Matching algorithm (Kuhn, 1955), computing semantic similarity of record contents. This ensures that each generated record matches the most appropriate reference. After alignment of records, we measure cell-level correctness using two kinds of metrics: **string-based** and **quantity-based**.

String-based metrics have traditionally been used in this space (for an overview of typical usage, see Timmer et al. (2025)). Here, we use precision, recall, and F_1 to quantify the correctness of the extracted cell values. We also use a character-based overlap metric, ChrF (Popović, 2015) to measure the degree to which the generated answer differs from the ground-truth answer. ChrF is language- and tokenisation-independent, making it a good fit for scientific data extraction.

Quantity-based metrics have been used in particular disciplines (e.g., enzyme work (Boorla and Maranas, 2025)), but we note their general applicability in our unified STG task. Here, we use these metrics to measure the distance of extracted numerical values to the ground truth values, revealing insights about the appropriateness of the *form* of the value, for example, with respect to unit consistency and scale. For numerical fields, we compute coverage, which reflects the fraction of entries with a non-null prediction. We also calculate mean absolute error (MAE) to measure the average deviation of these predictions, and the proportion of predictions within one order-of-magnitude ($p1mag$) to capture scale-level accuracy. All metrics are computed only on available predictions.

Formulas for all metrics with record alignment are provided in Appendix B.

6 Gauging STG Task Difficulty

6.1 Method

To gauge the problem difficulty, we employed typical LLM approaches, text-based and vision-based LLMs. Our goal is to use these approaches to characterise the challenges of the STG task through analyses of the limitations in applying generative AI to this task.

For the former, we used both closed-source and open-source models. For text-based generation, we used examples of locally-hostable models, GPT-OSS 120B (OpenAI et al., 2025), LLaMA-3.3-70B-Instruct (Grattafiori et al., 2024), as well as cloud-hosted models, such as o1-mini (Jaech et al., 2024).[6] These models operate on text

[6]We used the `gpt-o1-mini-2024-05-13` model via the

parsed from PDFs. For vision-based generation, we use Nanonets-OCR-s (Mandal et al., 2025)[7], a lightweight 3B vision-language model based on Qwen2.5-VL-3B-Instruct, designed for semantic understanding of document images.

For the LLM models, we follow the IE pipeline of Ghosh et al. (2024)[8], which integrates LangChain with GROBID (Lopez, 2009) for PDF parsing.[9] We note that GROBID is a widely used tool for PDF parsing, especially for scientific literature (Ghosh et al., 2024; Timmer et al., 2025).[10]

For the vision-language model, we provide each page as an image. All approaches use the same instruction to extract relevant information, given the respective inputs.[11]

6.2 Analysis

Table 2 summarises the overall performance of the sampled state-of-the-art approaches on the benchmark datasets. Note that the scores are calculated using our introduced *tuple-based* protocol, which first aligns predicted records with ground truth records. With this initial filtering, our proposed evaluation is stricter than cell-based evaluation approaches that perform cell-based evaluation (for an overview of such evaluation strategies, see Timmer et al. (2025)). This strictness reflects the constraint that the correctness of cell values must be judged in the context of the record, aligning with scientists' expectations of correctness. Our immediate observation is that the absolute values for the scores are quite low. ChrF, which can be interpreted as a percentage of matched character sequences, has a low average score of 7.74 (o1-mini). Similarly, the average F_1 is similarly low (e.g., o1-mini:13.78). This is true regardless of which approach we use, indicating the overall task difficulty and showing that current AI methods do not currently do well with the record-level focus of this task.

PDF Conversion Issues A limitation of the system lies in parsing PDFs. While GROBID is commonly used in the scientific domain (Ghosh et al., 2024; Timmer et al., 2025), it struggles with scanned documents, unusual section headers and embedded figures, often resulting in missing or misaligned content. Other OCR-based tools, such as Mathpix[12] and PyMuPDF[13], have shown potential to handle issues (Jiang et al., 2025). Here, we evaluated other OCR-based PDF conversion techniques but found GROBID to perform the conversion best.[14] We note that even web-based access to LLMs that allow PDF uploads, such as GPT-5 (OpenAI et al., 2025), rely on similar underlying logic for parsing and are subject to the same limitations.

LLM and VLM Context Limitations A key limitation of LLMs lies in their output token constraints (o1-mini: 4096 tokens)[15]. Datasets that have documents with multiple records (e.g., LegoBench or BRENDA-enzyme) often exceed this limit. While chunking mitigates this issue partially, it introduces new problems, as information is often spread throughout a paper, and splitting the document leads to incomplete or inconsistent outputs.

We observe that the evaluated VLM baseline (page-level extraction) performs worse than the evaluated text-based LLMs (with PDF parsing preprocessing). However, this comparison is not fully controlled, and two factors besides input modality may contribute to the gap. First, although Nanonets is designed for document understanding, it processes inputs at the page level. In our setting, accurate table generation often requires aggregating evidence spread across multiple pages, so page-level processing is a likely source of the gap. Second, Nanonets (3B parameters) is substantially smaller than several of the text models (e.g., GPT-OSS and LLaMA 3.3), and model scale may account for part of the difference. Our results are therefore consistent with the hypothesis that methods operating over the full document, abstracting away from presentation artefacts such as page boundaries, are better suited to STG, as suggested by the stronger performance of o1-mini, GPT-OSS and LLaMA 3.3. Disentangling the effects of modality, context scope and model scale would require controlled comparisons, for example a multi-page VLM of comparable size, or text models restricted to page-

Azure OpenAI API with options for data privacy protection.

[7] https://huggingface.co/nanonets/Nanonets-OCR-s

[8] https://github.com/SatanuG/ad_hoc_information_extraction

[9] Details of experimental settings are provided in Appendix C.1.

[10] We also tested alternative PDF parsing approaches, including vision-language models, but found that they did not perform as well as GROBID (Appendix C.2).

[11] See Appendix C for details of the prompt.

[12] https://mathpix.com

[13] https://pymupdf.readthedocs.io

[14] See Appendix D for details.

[15] The context windows and output budgets of recent models are steadily increasing, which may reduce the need for chunking.

level input.

Multi-record Issues LLMs can also overgenerate records per document. For instance, in the FE dataset, where each document corresponds to a single gold record (see Table 1), the o1-mini generated 1,738 records for just 326 documents. Although this boosts recall, it lowers precision and introduces evaluation misalignment (Ghosh et al., 2024). As noted above, since VLMs such as Nanonets operate on single pages, they lack the document-level context needed for accurate multi-record extraction.

Information Specificity and Source Fidelity We observe that systems sometimes extract information with extra detail that is not present in the gold records. For example, in the PPE dataset, the system extracts "PTzBT-14HD" instead of "PTzBT-14HD (Mn = 73 kDa)". Similarly, LLMs were observed to erroneously append "%" to a numeric value. Providing few-shot exemplars in the prompt that follow the expected schema may help to reduce such formatting issues (Ghosh et al., 2024; Bölücü et al., 2025). Another solution is to train models on task-specific data to ensure consistent adherence to the schema (Sundar et al., 2024; Dagdelen et al., 2024; Wei et al., 2025).

Schema Size and Performance Figure 3 shows the relationship between dataset schema size and F_1 score for the o1-mini model across 20 scientific datasets. The analysis indicates a negligible correlation (slope= -0.29, p= 0.50) between the number of columns (schema size) and performance. This suggests that the difficulty of the IE task is due to the inherent nature of the datasets and challenges related to the discipline-specific science. Datasets from similar domains do seem to cluster together, with Material Science and Chemistry datasets generally performing well, while Machine Learning datasets tend to have lower F_1 scores.

Differences Across Domains Figure 4 shows radar plots of model performance with string-based metrics averaged at the domain level. Due to space constraints, we show graphs for two scientific domains, *Chemistry* and *Material Science*, for which we have multiple datasets.[16] The results reveal a consistent pattern: all models achieve higher recall but lower precision. This is consistent with our observation that the evaluated models tend to *overgenerate* records, which explains the tendency for

[16]For graphs of all domains, see Appendix C.3.

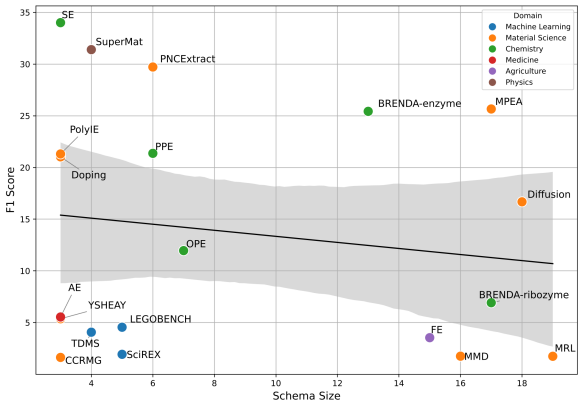
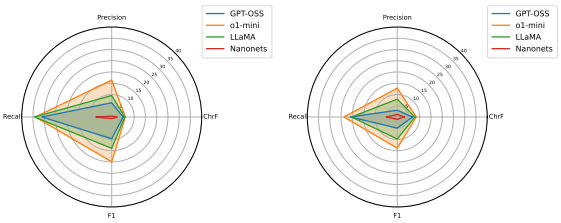


Figure 3: Correlation between F_1 score and schema size for the o1-mini model across 20 datasets.

recall to be high. Recall is also influenced by the density of records in each dataset. Datasets with fewer records per document tend to have higher recall, as models generate more candidate records.



(a) Chemistry (b) Material Science

Figure 4: Radar plots showing the performance of GPT-OSS, o1-mini, LLaMA 3.3, and Nanonets (vision) across domains. Metrics include ChrF, Precision, Recall, and F_1, averaged over all datasets in the domains.

Quantity-based Performance Assessments Our analysis of numerical values shows that this field is challenging. We note that the results show that extraction performance varies widely across domains and datasets. Coverage often remains low in complex material science and agriculture datasets, reflecting the difficulty of predicting all reference values. MAE values can be high, particularly in datasets with large numeric ranges (e.g., LEGOBENCH or BRENDA-enzyme), indicating substantial deviations from reference values when predictions exist. We suspect that this might be due to the extraction of values in different units of measurement The $p1mag$ metric, which indicates the percentage of values within one order of magnitude, is frequently high, suggesting that even when exact numerical predictions are inaccurate, they often remain within an order-of-magnitude. Certain

Chemistry datasets, such as BRENDA-ribozyme and PPE, show better alignment across metrics, likely reflecting more standardised reporting of numerical attributes. These results underscore the challenges of extracting precise numerical data across diverse scientific domains and highlight the need for improved normalisation, alignment, and coverage strategies in future work. For the full results, see Table 5 in Appendix C.4.

7 STG Challenges

Generating scientific tables from literature is challenging due to diverse data modalities, terminologies, and reporting styles. At a high-level we note that the task is diverse. For example, scientific domains differ in required output granularity: some datasets expect a single summary record per paper (e.g., FE), while others need many fine-grained records (e.g., MPEA, Diffusion).

As with any dataset, there are data issues. We also observe empty or missing values in annotations, sometimes deliberately left blank, other times due to annotation errors. In some cases, models correctly extract values that do not appear in the gold data, highlighting flaws in reference annotations (i.e., representation inconsistencies).

In addition to the issues highlighted in the analysis of Section 6, we highlight key open challenges that may accelerate the process in automated table generation.[17]

Representing Numerical Values Evaluation of text variants remains an open problem as exact-match metrics penalise these semantically equivalent but superficially different results. We frequently observe inconsistencies between predicted and gold values for quantitative attributes (e.g., temperature, kcat).

For example, we observed **unit mismatches** where values are expressed using different units than the gold data (e.g., "10 kbar" instead of "1 GPa"), leading to false negatives under exact-match evaluation unless unit normalisation is applied. The other observation is **representation inconsistencies** where there are variations in formatting (e.g., "2–2.5 GPa" versus "2 to 2.5 GPa") which result in semantically equivalent values but still fail exact-match checks. This highlights a common surface-form mismatch where the content is correct, but the formatting differs. These kinds of mismatches can be resolved through normalisation. Other examples include alternative ways to report numerical precision. (e.g., "80 wt%" vs. "0.8"; "2.5" vs. "2.50"; "1.69e-11" vs. "0.0000000000169"). Alternations are numerically equivalent but expressed differently. This highlights the need for normalisation to ensure an accurate evaluation.

Unit and Quantity Standardisation Scientific texts frequently report quantitative values in diverse units, making it challenging to compare and integrate data from different studies.

For example, in enzyme kinetics, the Michaelis constant (K_m) can be reported in various units, such as micromolar (μM), molar (M), or millimolar (mM), each of which requires unit conversion for consistency. Moreover, human-curated databases like BRENDA (Chang et al., 2021) and ChEMBL (Papadatos et al., 2015) standardise units by normalising values across different measurements. Such unit standardisation ensures that data can be easily compared and integrated for downstream analyses and building knowledge-bases (e.g., BRENDA). Using an existing library for normalising units of measurement (the pint library) demonstrated an improvement in performance for the BRENDA-enzyme dataset (F_1 34.2 $\rightarrow$ 38.7), demonstrating the potential for further research to improve STG performance more widely.[18]

Entity Normalisation In scientific literature, the same entity is often represented in multiple forms. For example, "TiO_2", "titania", "AO_2 (A = Ti)", and "titanium dioxide" all refer to the same specific stoichiometry: "TiO_2". To ensure consistency and improve searchability, it is important to normalise these entities into a standard, canonical format. This enables, for example, a query for "titania" to also retrieve papers that mention "titanium dioxide".

In preliminary trials to include some known value normalisation techniques, we observed a performance gain (F_1 increasing: $3.59 \rightarrow 18.56$) for datasets from machine learning. This suggests an opportunity for similar techniques to be developed for other subsets of the benchmark dataset.

Entity Linking In scientific literature, entities such as genes, proteins, and chemicals are often referenced using different names and formats, which can make it difficult to integrate and compare information across studies. The task of entity linking

[17] Descriptions of preliminary attempts to tackle these challenges are described in Appendix D.

[18] https://github.com/hgrecco/pint

addresses this challenge by mapping these entities to their corresponding entries in external structured knowledge bases (e.g., Gene Ontology (GO) (Ashburner et al., 2000), UniProt (Bairoch et al., 2005), and PubChem (Wang et al., 2009)). This ensures that multiple references to the same entity are consistently identified, improving data interoperability and enabling more accurate analysis of scientific findings across sources. In the literature, this task is also referred to as table entity linking (TEL), which specifically aims to map entity mentions in tables to their corresponding entities in a knowledge base (Lou et al., 2023; Wu et al., 2024; Wang et al., 2024c). We leave explorations of entity linking to future work, noting that this is a challenging scenario in which the linked category may not occur verbatim in the source text.

Table and Figure Understanding Scientific papers often present key findings in tables and figures, which require different strategies for extraction. Table understanding involves detecting regions, recognising structures (e.g., headers, merged cells, hierarchies), and parsing the content into a machine-readable format. Table designs vary considerably across domains, from simple grids to complex, nested layouts with implicit row or column relationships (Sun and Liu, 2015; Moosavi et al., 2021; Darmanović et al., 2023).

Figure understanding, on the other hand, focuses on interpreting visual elements such as bar charts, scatter plots, and line graphs. While some figures explicitly encode data (e.g., ChartQA (Masry et al., 2022; Kim et al., 2025)), others convey insights more implicitly through visual trends and relative positions (e.g., Diffusion and MMD datasets). Effectively utilising tables and figures for tasks requires a deep understanding of both visual structure and contextual semantics across modalities. It shows that integrating multimodality helps to extract more accurate information from documents (Khalighinejad et al., 2025). Again, we leave a treatment of tables and figures to future work.

8 Conclusion

We introduce *Scientific Table Generation* (STG), a new task and benchmark designed to unify diverse structured information extraction efforts within a table-centric framework. The benchmark combines 20 datasets from six scientific domains, providing a rich and diverse testbed for evaluating systems that aim to structure scientific content based on predefined schemas. Our formalisation of the STG task, curation of a benchmark dataset, and introduction of a record-based evaluation protocol highlight the difficulty of the task and the limitations of standard zero-shot pipelines for this task.

Our analysis of the output of typical LLM and VLM approaches to document processing reveals several ongoing challenges in the field, including inconsistencies in how values are reported, variations in entity mentions, lack of standard units in extractions, PDF conversion issues, and gaps between text, tables, and figures. All emerge as promising areas for further research.

To promote progress in this area, we make our benchmark and codebase publicly available and invite the research community to contribute new datasets across additional scientific domains. This collaborative approach will help expand STG and enable broader development of models that not only extract information but also structure and normalise it in ways that align with real-world scientific workflows.

Limitations

The study provides a unified and valuable benchmark for the scientific table generation (STG) task. It covers six domains (machine learning, material science, chemistry, agriculture, medicine, and physics). However, there are several limitations. First, the current benchmark is constrained to these six domains; expanding to additional areas such as engineering or environmental science could reveal new domain-specific challenges and improve generalisability.

Second, although the benchmark integrates 20 datasets under a common task formulation, each dataset has its own specialised schema and varies in granularity (coarse- or fine-grained). This variation limits the comparability of results across datasets. Moreover, the dataset sizes are relatively small, which may limit the ability to fine-tune models specifically for each dataset.

Third, we propose subtasks and provide simple baseline solutions to highlight their utility. However, it requires more research to systematically evaluate the role of these subtasks and to explore their integration into multimodal pipelines.

Finally, we do not re-implement or run the methods proposed in the original studies from which the datasets are drawn. As such, our focus remains on

providing a unified benchmark and evaluating general LLM and VLM-based baselines, rather than directly comparing against domain-specific models reported in prior work.

Ethics Statement

Our benchmark is built on existing public datasets adopted from existing repositories. Therefore, we do not foresee any serious or harmful issues related to their content. To ensure full compliance with copyright laws, we provide only the necessary identifiers (e.g., DOI, PubMed ID) for the scientific documents, but we do not distribute the actual documents. Researchers are required to download the relevant files independently. The repository includes detailed instructions to guide users on how to properly place the downloaded documents.

References

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, and 4 others. 2018. Construction of the Literature Graph in Semantic Scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics.

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*.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, and 1 others. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics.

Amos Bairoch, Rolf Apweiler, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, and 1 others. 2005. The universal protein resource (UniProt). *Nucleic acids research*, 33(suppl_1):D154–D159.

Necva Bölücü, Jessica Irons, Changhyun Lee, Brian Jin, Maciej Rybinski, Huichen Yang, Andreas Duenser, and Stephen Wan. 2026. Using a human-AI teaming approach to create and curate scientific datasets with the SciLire system. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 428–444, Rabat, Marocco. Association for Computational Linguistics.

Necva Bölücü, Jordan Pennells, Huichen Yang, Maciej Rybinski, and Stephen Wan. 2025. An Evaluation of Large Language Models for Supplementing a Food Extrusion Dataset. *Foods*, 14(8):1355.

Veda Sheersh Boorla and Costas D. Maranas. 2025. CatPred: a comprehensive framework for deep learning in vitro enzyme kinetic parameters. *Nature Communications*, 16(1).

Rohit Borah, Andrew W Brown, Patrice L Capers, and Kathryn A Kaiser. 2017. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ open*, 7(2):e012545.

Christopher KH Borg, Carolina Frey, Jasper Moh, Tresa M Pollock, Stéphane Gorsse, Daniel B Miracle, Oleg N Senkov, Bryce Meredig, and James E Saal. 2020. Expanded dataset of mechanical properties and observed phases of multi-principal element alloys. *Scientific Data*, 7(1):430.

Ekaterina Borisova, Nikolas Rauscher, and Georg Rehm. 2025. SciVQA 2025: Overview of the First Scientific Visual Question Answering Shared Task. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 182–210, Vienna, Austria. Association for Computational Linguistics.

Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Wang Changxin, Zhifeng Gao, Hongshuai Wang, Li Yongge, Mujie Lin, Shuwen Yang, Jiankun Wang, Mingjun Xu, Jin Huang, Xi Fang, Jiaxi Zhuang, Yuqi Yin, Yaqi Li, Changhong Chen, and 4 others. 2025. SciAssess: Benchmarking LLM Proficiency in Scientific Literature Analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2335–2357, Albuquerque, New Mexico. Association for Computational Linguistics.

Hengxing Cai, Xiaochen Cai, Shuwen Yang, Jiankun Wang, Lin Yao, Zhifeng Gao, Junhan Chang, Sihang Li, Mingjun Xu, Changxin Wang, and 1 others. 2024. Uni-SMART: Universal Science Multimodal Analysis and Research Transformer. *arXiv preprint arXiv:2403.10301*.

Antje Chang, Lisa Jeske, Sandra Ulbrich, Julia Hofmann, Julia Koblitz, Ida Schomburg, Meina

Neumann-Schaal, Dieter Jahn, and Dietmar Schomburg. 2021. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic acids research*, 49(D1):D498–D508.

Hong Chen, Hiroya Takamura, and Hideki Nakayama. 2021. SciXGen: A scientific paper dataset for context-aware text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1483–1492, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jerry Cheung, Yuchen Zhuang, Yinghao Li, Pranav Shetty, Wantian Zhao, Sanjeev Grampurohit, Rampi Ramprasad, and Chao Zhang. 2024. POLYIE: A Dataset of Information Extraction from Polymer Material Scientific Literature. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2370–2385, Mexico City, Mexico. Association for Computational Linguistics.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Filip Darmanović, Allan Hanbury, and Markus Zlabinger. 2023. SCI-3000: A Dataset for Figure, Table and Caption Extraction from Scientific PDFs. In *International Conference on Document Analysis and Recognition*, pages 234–251. Springer.

Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. 2022. Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv:2212.05238*.

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*.

Luca Foppiano, Sae Dieb, Akira Suzuki, Pedro Baptista de Castro, Suguru Iwasaki, Azusa Uzuki, Miren Garbine Esparza Echevarria, Yan Meng, Kensei Terashima, Laurent Romary, Yoshihiko Takano, and Masashi Ishii. 2021. SuperMat: construction of a linked annotated dataset from superconductors-related publications. *Science and Technology of Advanced Materials: Methods*, 1(1):34–44.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. 2025. JSONSchemaBench: A Rigorous Benchmark of Structured Outputs for Language Models. *Preprint*, arXiv:2501.10868.

Satanu Ghosh, Neal Brodnik, Carolina Frey, Collin Holgate, Tresa Pollock, Samantha Daly, and Samuel Carton. 2024. Toward Reliable Ad-hoc Scientific Information Extraction: A Case Study on Two Materials Dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15109–15123, Bangkok, Thailand. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sameera Horawalavithana, Ellyn Ayton, Shivam Sharma, Scott Howland, Megha Subramanian, Scott Vasquez, Robin Cosbey, Maria Glenski, and Svitlana Volkova. 2022. Foundation Models of Scientific Knowledge for Chemistry: Opportunities, Challenges and Lessons Learned. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 160–172, virtual+Dublin. Association for Computational Linguistics.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2019. Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5203–5213, Florence, Italy. Association for Computational Linguistics.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation Extraction By End-to-end Language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A Challenge Dataset for Document-Level Information Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-Level N-ary Relation Extraction with Multiscale Representation Learning. In *Proceedings of*

the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3693–3704, Minneapolis, Minnesota. Association for Computational Linguistics.

Jinling Jiang, Jie Hu, Siwei Xie, Menghao Guo, Yuhang Dong, Shuai Fu, Xianyue Jiang, Zhenlei Yue, Junchao Shi, Xiaoyu Zhang, and 1 others. 2025. Enzyme Co-Scientist: Harnessing Large Language Models for Enzyme Kinetic Data Extraction from Literature. *bioRxiv*, pages 2025–03.

Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji, and Jiawei Han. 2023. Instruct and Extract: Instruction Tuning for On-Demand Information Extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10030–10051, Singapore. Association for Computational Linguistics.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Ilknur Karadeniz and Arzucan Özgür. 2019. Linking entities through an ontology using word embeddings and syntactic re-ranking. *BMC bioinformatics*, 20(1):156.

Ghazal Khalighinejad, Defne Circi, L. Brinson, and Bhuwan Dhingra. 2024. Extracting Polymer Nanocomposite Samples from Full-Length Documents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13163–13175, Bangkok, Thailand. Association for Computational Linguistics.

Ghazal Khalighinejad, Sharon Scott, Ollie Liu, Kelly L. Anderson, Rickard Stureborg, Aman Tyagi, and Bhuwan Dhingra. 2025. MatViX: Multimodal Information Extraction from Visually Rich Articles. 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 3636–3655, Albuquerque, New Mexico. Association for Computational Linguistics.

Wonjoong Kim, Sangwu Park, Yeonjun In, Seokwon Han, and Chanyoung Park. 2025. SIMPLOT: Enhancing Chart Question Answering by Distilling Essentials. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 573–593, Albuquerque, New Mexico. Association for Computational Linguistics.

David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Adithya Kulkarni, Fatimah Alotaibi, Xinyue Zeng, Longfeng Wu, Tong Zeng, Barry Menglong Yao, Minqian Liu, Shuaicheng Zhang, Lifu Huang, and Dawei Zhou. 2025. Scientific Hypothesis Generation and Validation: Methods, Datasets, and Future Directions. *Preprint*, arXiv:2505.04651.

Feiran Li, Le Yuan, Hongzhong Lu, Gang Li, Yu Chen, Martin KM Engqvist, Eduard J Kerkhoven, and Jens Nielsen. 2022. Deep learning-based k cat prediction enables improved enzyme-constrained model reconstruction. *Nature Catalysis*, 5(8):662–672.

Patrice Lopez. 2009. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Corfu, Greece, September 27-October 2, 2009. Proceedings 13*, pages 473–474. Springer.

Yuze Lou, Bailey Kuehl, Erin Bransom, Sergey Feldman, Aakanksha Naik, and Doug Downey. 2023. S2abEL: A Dataset for Entity Linking from Scientific Tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3089–3101.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Xinyuan Lu, Liangming Pan, Yubo Ma, Preslav Nakov, and Min-Yen Kan. 2025. TART: An Open-Source Tool-Augmented Framework for Explainable Table-based Reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4323–4339, Albuquerque, New Mexico. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. BioRED: A Rich Biomedical Relation Extraction Dataset. *Briefing in Bioinformatics*.

Souvik Mandal, Ashish Talewar, Paras Ahuja, and Prathamesh Juvatkar. 2025. Nanonets-OCR-S: A model for transforming documents into structured markdown with intelligent content recognition and semantic tagging.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. Scigen: a dataset for reasoning-aware text generation from scientific tables. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Ahmed Nassar, Nikolaos Livathinos, Maksym Lysak, and Peter Staar. 2022. TableFormer: Table Structure Understanding With Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4614–4623.

Benjamin Newman, Yoonjoo Lee, Aakanksha Naik, Pao Siangliulue, Raymond Fok, Juho Kim, Daniel S Weld, Joseph Chee Chang, and Kyle Lo. 2024. ArxivDIGESTables: Synthesizing Scientific Literature into Tables using Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9612–9631, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b gpt-oss-20b Model Card. *Preprint*, arXiv:2508.10925.

George Papadatos, Anna Gaulton, Anne Hersey, and John P Overington. 2015. Activity, assay and target data curation and quality in the ChEMBL database. *Journal of computer-aided molecular design*, 29:885–896.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter W J Staar. 2022. DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis.

Maciej P Polak and Dane Morgan. 2024. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1):1569.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Chandan K Reddy and Parshin Shojaee. 2025. Towards scientific discovery with generative ai: Progress, opportunities, and challenges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28601–28609.

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

Shruti Singh, Shoaib Alam, Husain Malwat, and Mayank Singh. 2024. LEGOBench: Scientific Leaderboard Generation Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14598–14613, Miami, Florida, USA. Association for Computational Linguistics.

Yu Song, Santiago Miret, and Bang Liu. 2023. MatSciNLP: Evaluating Scientific Language Models on Materials Science Language Tasks Using Text-to-Schema Modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3621–3639, Toronto, Canada. Association for Computational Linguistics.

Haoyang Su, Renqi Chen, Shixiang Tang, Xinzhe Zheng, Jingzhe Li, Zhenfei Yin, Wanli Ouyang, and Nanqing Dong. 2024. Two Heads Are Better Than One: A Multi-Agent System Has the Potential to Improve Scientific Idea Generation. *arXiv preprint arXiv:2410.09403*.

Lijun Sun and Yao Liu. 2015. Review of Research on Table Extraction in Scientific and Technical Literature. *International Journal of Knowledge and Language Processing*, 6:48–62.

Anirudh Sundar, Christopher Richardson, and Larry Heck. 2024. gtbls: Generating tables from text by conditional question answering. *arXiv preprint arXiv:2403.14457*.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical Entity Representations with Synonym Marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

Deep Search Team. 2024. Docling Technical Report. Technical report.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards Domain-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore. Association for Computational Linguistics.

Roelien C. Timmer, Yufang Hou, and Stephen Wan. 2025. A Position Paper on the Automatic Generation of Machine Learning Leaderboards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30749–30772, Suzhou, China. Association for Computational Linguistics.

Benjamin Townsend, Eamon Ito-Fisher, Lily Zhang, and Madison May. 2021. Doc2dict: Information extraction as text generation. *arXiv preprint arXiv:2105.07510*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and

Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. 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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 50622–50649. PMLR.

Xingbo Wang, Samantha L Huey, Rui Sheng, Saurabh Mehta, and Fei Wang. 2024b. SciDaSynth: Interactive Structured Knowledge Extraction and Synthesis from Scientific Literature with Large Language Model. *arXiv preprint arXiv:2404.13765*.

Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, and Stephen H Bryant. 2009. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research*, 37(suppl_2):W623–W633.

Yuanzheng Wang, Yixing Fan, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. 2024c. RoCEL: Advancing Table Entity Linking through Distinctive Row and Column Contexts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15284–15298, Miami, Florida, USA. Association for Computational Linguistics.

Galen Wei, Xinchun Ran, Runeem AI-Abssi, and Zhongyue Yang. 2025. Finding the dark matter: Large language model-based enzyme kinetic data extractor and its validation. *Protein Science*, 34(9):e70251.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, and 1 others. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Tianxing Wu, Lin Li, Huan Gao, Guilin Qi, Yuxiang Wang, and Yuehua Li. 2024. Npel: Neural paired entity linking in web tables. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Tong Xie, Yuwei Wan, Wei Huang, Zhenyu Yin, Yixuan Liu, Shaozhou Wang, Qingyuan Linghu, Chunyu Kit, Clara Grazian, Wenjie Zhang, and 1 others. 2023a. Darwin series: Domain specific large language models for natural science. *arXiv preprint arXiv:2308.13565*.

Tong Xie, Yuwei Wan, Wei Huang, Yufei Zhou, Yixuan Liu, Qingyuan Linghu, Shaozhou Wang, Chunyu Kit, Clara Grazian, Wenjie Zhang, and 1 others. 2023b. Large language models as master key: unlocking the secrets of materials science with GPT. *arXiv preprint arXiv:2304.02213*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. DWIE: An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*, 58(4):102563.

Youxue Zhang, Huaiwei Ni, and Yang Chen. 2010. Diffusion data in silicate melts. *Reviews in Mineralogy and Geochemistry*, 72(1):311–408.

A Benchmark Insights

We present the details of 6 scientific domains with the 20 datasets used in the benchmark. We summarise the source, content focus, and key characteristics of each dataset.

A.1 Datasets

Machine Learning The machine learning domain focuses on developing and evaluating models across various tasks, often through benchmark datasets (Hou et al., 2019; Singh et al., 2024). Consequently, leaderboard generation becomes important for extracting relevant information and comparing model performance effectively, as well as for table generation from scientific documents. In this domain, we collect different-scale datasets focused on leaderboard generation.

TDMS is a scientific leaderboard dataset that contains information on tasks, datasets, evaluation metrics, and scores from NLP papers. It targets the extraction of the best numerical results from experimental sections.

SciREX supports scientific information extraction, specifically relation extraction and entity linking, based on material, method, and evaluation metrics.

LegoBench is a dataset designed for leaderboard generation, reformulated as a question-answering

Dataset	Schema	Source
TDMS	Task, Dataset, Metric, Score	ACL Anthology ID
SciREX	Material, Method, Metric, Task, Score	Semantic Scholar ID
LEGOBENCH	Task, Dataset, Method, Metric, Score	ArXiv ID
MPEA	high entropy alloy formula, microstructure, processing method, BCC/FCC/other, grain size, experimental density, hardness, type of test, test temperature, yield strength, ultimate tensile strength, elongation, elongation plastic, experimental young modulus, oxygen content, nitrogen content, carbon content	DOI
Diffusion	melt, diffusing species, type of diffusion, type of experiment, test temperature, pressure, diffusivity, SiO2, TiO2, Al2O3, FeOt, MnO, MgO, CaO, Na2O, K2O, P2O5, H2Ot	DOI
YSHEAY	Material, Value, Unit	DOI
CCRMG	Material, Value, Unit	DOI
Doping	BASEMAT, DOPANT, DOPMODQ	DOI
MMD	Formula, Formula units per cell, Atomic sites per cell, Crystal system, Formation energy (eV/atom), Energy relative to convex hull (eV/atom), Structure search, Averaged magnetic moment (μB/atom), Magnetic polarization, Js</(T), Magnetic easy axis, Magnetic anisotropy constants:Ka-c, Magnetic anisotropy constants:Kb-c (MJ/m3), Magnetic anisotropy constants:Kb-a (MJ/m3), Magnetic anisotropy constants:Kd-a (MJ/m3), Curie temperature, TC(K), Methods	DOI
MRL	Electrical resistivity (Ωcm), Seebeck coefficient (μCV/K), ZT (thermoelectric figure of merit), formula, synthesis, form, temperature, Thermal conductivity (W/mK), Average atomic mass (g/mol), Scarcity (wt fraction/abundance), HHIP (Herfindahl-Hirschman Index for elemental production), HHIR (Herfindahl-Hirschman Index for elemental reserves), Unit cell volume (Å^3), Average atomic volume (Å^3), Atoms per unit cell, Electrical conductivity (Ω^{-1} cm^{-1}), Power Factor (W/(K^2m)), S^2 (μV/K)^2, Electronic fraction of thermal conductivity (κ_e/κ)	DOI
PolyIE	Chemical Name, Property Name, Property Value	DOI
PNCExtract	Matrix Chemical Name, Matrix Abbreviation, Filler Chemical Name, Filler Abbreviation, Filler Composition Mass, Filler Composition Volume	DOI
BRENDA-enzyme	Enzyme, Organism, Substrate, km, unit_km, kcat, unit_kcat, km_kcat, unit_km_kcat, Commentary[Temp], Commentary[pH], Commentary[Mutant], Commentary[Cosubstrate]	PubMed
BRENDA-ribozyme	Ribozyme, Organism, Variable_reactant, Kobs, Unit_Kobs, km, Unit_Km, kcat, Unit_Kcat, km_kcat, Unit_Kcat/Km, Kcleav, Unit_Kcleav, Commentary[Temp], Commentary[pH], Commentary[Mutant], Commentary[Others]	PubMed
OPE	Host, Dopant, Td [$^\circ$ C]/Tg [$^\circ$ C]/ET [eV], Von [V], EQE, CE, PE, CIE [x, y]	DOI
PPE	Nickname, PCE_max(%), PCE_ave(%), Voc (V), Jsc (mA cm^2), FF	DOI
SE	solute_name, solvent_name, temperature, pressure, solubility	DOI
AE	Compound, Affinities	DOI
FE	extruder product type, extruder manufacturer company, extruder model name, extruder manufacturer country, scale, screw type, screw dimension, length to diameter ratio, die dimensions, ingredients, moisture in extrusion trials, input variables, response variables, feed characterisation, quality characterisation	DOI
SuperMat	Material, tcValue, Pressure, me_method	DOI

Table 3: Overview of 20 datasets across six domains. Each dataset is linked to a document source and schema.

(QA) task over documents, focusing on extracting structured records from full-text papers. We converted the QA dataset into the STG formalised dataset.

Material Science Materials science explores a wide variety of substances, including polymers, ceramics, and metals. Each of these has distinct physical, chemical, and mechanical properties, making them important for applications in industries such as aerospace, electronics, and construction. Scientific literature in this domain often reports not only material compositions and target properties, but also detailed experimental and contextual information. For example, the MPEA dataset includes parameters such as melt composition, pressure, temperature, and diffusion type. Materials science is a well-studied domain with many datasets covering

detailed information found in the literature (Borg et al., 2020; Khalighinejad et al., 2025).

MPEA is a dataset focused on multi-principal element alloys (MPEAs), also known as high-entropy alloys (HEAs). It contains experimental key properties (e.g. strength, hardness, and Young's modulus).

Diffusion is a dataset that covers the diffusion (transport of elements) in silicate melts and glasses, which is a key process in both geology and materials science. It captures experimental parameters such as temperature, pressure, and diffusion coefficients. This dataset is built on a summary table presented by Zhang et al. (2010). Instead of extracting information from this paper, we gathered the reference papers to build a dataset for this task similar to Ghosh et al. (2024). Similar to this dataset, there are more possible datasets that we have not included.

YSHEAY and CCRMG cover the yield strength of HEAs and the critical cooling rates of metallic glasses, respectively. Both YSHEAY and CCRMG datasets include material names, measured property values, and units. We formulate datasets and gather referenced papers given in the datasets for our benchmark. The datasets extract the sentences from the scientific documents in which the information is presented to simplify the task. We use the full document, similar to the rest of the other datasets.

MRL is derived from a review on thermoelectric materials and includes experimental results and properties (e.g., Electrical resistivity, Seebeck coefficient, thermoelectric figure of merit).

Doping is a dataset for materials chemistry that covers dopant (also called a doping agent), which is a small amount of a substance added to a material to alter its physical properties, such as electrical or optical properties, with its host materials.

MMD is a dataset focused on physical and magnetic properties of materials relevant to magnetism research and materials design including physical and magnetic properties of materials relevant to magnetism research and materials design.

PolyIE is a dataset for polymer scientific literature covering four different material science domains: polymer solar cells, ring-opening polymerization, polymer membranes, and lithiumion batteries. We could not access all documents in the dataset since the DOI numbers are not provided.

PNCExtract is a dataset focused on the extraction of polymer nanocomposites (PNCs) samples from science articles. The variety in PNCs comes from different matrix and filler combinations that modify the properties.

Chemistry The field of chemistry involves a vast array of complex molecular structures, chemical reactions, and properties, alongside a substantial amount of data presented in formulas, reaction equations, and diagrams. Effectively processing and interpreting these components is a significant challenge for language models. Tasks in the chemical domain demand precise understanding of molecular compositions, reaction mechanisms, and material properties.

BRENDA is a database on functional and molecular information of enzymes, based on primary literature. Jiang et al. (2025) construct 2 datasets (enzyme and ribozyme) for enzyme and ribozyme kinetics data collected from scientific literature.

OPE is a dataset that covers OLED molecules and their optical properties. The dataset is built on summary tables from a limited number of scientific documents. Instead of that, we collect reference studies instead of using given documents to make it similar to other datasets in this study and remove (a text-based format for representing chemical structures as a line notation) from the schema since this attribute is not provided in the scientific documents.

PPE is a dataset that contains values like power conversion efficiency (PCE) and open-circuit voltage (VOC) for polymer molecules extracted from scientific documents.

Solubility Extraction (SE) is a dataset Organic electrolytes, extensively used in battery technologies, comprise organic solvents, lithium salts, and additives. Understanding solubility in organic electrolytes is crucial as it impacts the efficiency of electrolytic processes, product selectivity, and equipment design.

Medicine Medicine often involves complex interactions between biological molecules, experimental methods, and treatment conditions. Literature in this domain typically reports experimental results, molecular data, and pharmacological properties essential for therapeutic development.

Affinity Extraction (AE) extract an affinity table containing molecules' tags, SMILES, and their affinities to different targets in bioassays. It tests several key capabilities of LLMs, including understanding complex and domain-specific language, as well as molecules and tables. Affinity data extraction requires not just surface-level text processing

but also a deeper analysis to match different modalities.

Agriculture Agricultural research often involves complex interactions between biological materials, machinery, and environmental conditions. Literature in this domain typically reports experimental setups, crop properties, and processing parameters relevant to food production and sustainability.
Food Extrusion (FE) is a qualitative literature collection that consolidates and standardises research on food extrusion, including details about the extruder and other variables and characteristics explored in the articles

Physics Superconductors Material (SuperMat) is a manually linked annotated dataset of superconductor-related materials and properties.

A.2 Benchmark Curation and Cleaning

Source document recovery. We collect the original source documents for every dataset and rebuild each dataset from those documents, rather than reusing the released annotations as they are. However, for some datasets, we cannot find all the original documents. Some papers are not publicly available. Others are listed without a persistent identifier (e.g., DOI or ArXiv ID) or with too little bibliographic information to locate them reliably.[19] As a result, the number of datasets presented here may not exactly match the original resources. Finally, we do not follow the original dataset splits (train/dev/test), so we do not perform comparisons of baseline results with the methods proposed in the original papers.

Schema normalisation We convert the datasets into a unified table-centric format. We remove certain attributes when the corresponding values cannot be reliably extracted from the PDFs. One example is *SMILES* strings, which are rarely stated explicitly in the text and are usually derived from external databases. In contrast, we retain attributes whose values are expressed in the document but annotated in a normalised or canonical form (e.g., *Task* in TDMS and SciREX datasets), and we treat entity normalisation as one of the STG challenges.

A.3 Resources and Downloads

TDMS https://github.com/IBM/science-result-extractor

SciREX https://github.com/allenai/SciREX

LEGOBENCH https://github.com/lingo-iitgn/LEGOBench

MPEA https://github.com/CitrineInformatics/MPEA_dataset

Diffusion http://www.minsocam.org/MSA/RIM/rim72.html

YSHEAY and CCRMG https://figshare.com/articles/dataset/Data_for_Flexible_Model-Agnostic_Method_for_Materials_Data_Extraction_from_Text_Using_General_Purpose_Language_Models_/21861948

Doping https://github.com/lbnlp/NERRE

MRL and MMD https://github.com/MasterAI-EAM/Darwin

PolyIE https://github.com/jerry3027/PolyIE

PNCExtract https://github.com/ghazalkhalighinejad/PNCExtract

BRENDA (enzyme and ribozyme) https://github.com/JackKuo666/LLM-BioDataExtractor

OPE, PPE, SE, and AE https://github.com/sci-assess/SciAssess

FE https://data.csiro.au/collection/csiro%3A64834v1

SuperMat https://github.com/lfoppiano/SuperMat

B Evaluation metrics

Record Alignment The output table $T \in \mathcal{V}^{M \times K}$ is partitioned into N document-level subtables, where subtable $T^{(i)} \in \mathcal{V}^{R_i \times K}$ contains the R_i records contributed by document d_i. Correspondingly, the predicted table $\hat{T}$ is partitioned into subtables $\hat{T}^{(i)} \in \mathcal{V}^{\hat{R}_i \times K}$, where $\hat{R}_i$ may differ from R_i. Alignment is performed independently within each subtable pair $(T^{(i)}, \hat{T}^{(i)})$.

For each record, we concatenate its K attribute values into a single string and encode it with a sentence encoder ϕ. The pairwise semantic similarity matrix $\mathbf{S}^{(i)} \in \mathbb{R}^{R_i \times \hat{R}_i}$ is then computed as:

[19] Records whose source document could not be retrieved are excluded from the benchmark.

$$S^{(i)}_{m,\hat{m}} = \cos(\phi(g_m),\ \phi(\hat{g}_{\hat{m}})) \tag{1}$$

where $\phi(g_m)$ denotes the sentence embedding of the concatenated attribute values of record g_m, $\cos(\cdot,\cdot)$ denotes cosine similarity, and ϕ is implemented using `all-MiniLM-L6-v2`[20] (**?**).

Since $R_i \neq \hat{R}_i$ in general, $\mathbf{S}^{(i)}$ is a rectangular matrix. The Hungarian algorithm (Kuhn, 1955) finds a maximum-weight partial matching:

$$\pi_i^* = \arg\max_{\pi \in \Pi_i} \sum_{m=1}^{\min(R_i, \hat{R}_i)} S^{(i)}_{m,\pi(m)} \tag{2}$$

where Π_i is the set of all injective mappings from the smaller index set into the larger. This yields $\min(R_i, \hat{R}_i)$ matched pairs $\mathcal{M}_i = \{(g_m, \hat{g}_{\pi_i^*(m)})\}$, with the remaining $|R_i - \hat{R}_i|$ records left unmatched. Unmatched predicted records penalise precision; unmatched ground-truth records penalise recall. The global alignment set across all documents is $\mathcal{A} = \bigcup_i \mathcal{M}_i$.

Cell=level Scoring **String-based metrics** have traditionally been used in this space (for an overview of typical usage, see Timmer et al. (2025)). Here, we use precision, recall, and F_1 to quantify the correctness of the extracted cell values. We also use a character-based overlap metric, ChrF (Popović, 2015) to measure the degree to which the generated answer differs from the ground-truth answer. ChrF is language- and tokenisation-independent, making it a good fit for scientific data extraction.

Each record r is treated as a *bag of cells*: a matched pair $(r_j, \hat{r}_i) \in \mathcal{A}$ contributes $|\hat{r}_i \cap r_j|$ correctly predicted cells to the numerator, where a cell is counted as correct if its value exactly matches the ground-truth cell in the aligned column. Precision normalises by the total number of predicted cells across *all* predicted records (including unmatched ones), ensuring that hallucinated extra records are penalised:

$$\text{Precision} = \frac{\sum_{(i,j)\in\mathcal{A}} |\hat{r}_i \cap r_j|}{\sum_{\hat{r}\in\hat{\mathcal{R}}} |\hat{r}|}, \tag{3}$$

$$\text{Recall} = \frac{\sum_{(i,j)\in\mathcal{A}} |\hat{r}_i \cap r_j|}{\sum_{r\in\mathcal{R}} |r|}, \tag{4}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{5}$$

where $\hat{\mathcal{R}}$ and $\mathcal{R}$ denote the full sets of predicted and ground-truth records, respectively, and $|r|$ is the number of cells in record r.

We also report the micro-averaged ChrF score over aligned records:

$$\text{ChrF}_{\text{micro}} = \frac{\sum_{(i,j)\in\mathcal{A}} \sum_{c\,\in\,\text{cells}(r_j)} \text{ChrF}(\hat{c}_i,\ c)}{\sum_{(i,j)\in\mathcal{A}} |\text{cells}(r_j)|}, \tag{6}$$

where $\text{cells}(r_j)$ is the ordered set of cells in reference record r_j, $\hat{c}_i$ is the predicted cell in the corresponding column of $\hat{r}_i$, and $\text{ChrF}(\hat{c}_i, c)$ is the character-level F-score between the two cell strings. The denominator counts the total number of reference cells across all aligned pairs.

Quantity-based metrics have been used in particular disciplines (e.g., enzyme work (Boorla and Maranas, 2025)), but we note their general applicability in our unified STG task. Here, we use these metrics to measure the distance of extracted numerical values to the ground truth values, revealing insights about the appropriateness of the *form* of the value, for example, with respect to unit consistency and scale.

After record alignment, let N_f be the total number of numeric reference entries in column f across all aligned pairs $(\hat{r}_i, r_j) \in \mathcal{A}$, and let $A_f \subseteq \{1, \ldots, N_f\}$ be the index set of those entries for which a non-null numeric prediction is available. All metrics are computed only over A_f; non-numeric cells and missing predictions are excluded.

Coverage quantifies the proportion of numeric reference entries in column f for which a non-null prediction is available:

$$\text{Coverage}_f = \frac{|A_f|}{N_f} \times 100. \tag{7}$$

[20] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 Last visited: 01-04-2026

Mean Absolute Error (MAE) computes the average absolute deviation of available predictions from the reference values in column f:

$$\mathrm{MAE}_f = \frac{1}{|A_f|} \sum_{i \in A_f} |v_{f,i} - \hat{v}_{f,i}|, \tag{8}$$

where $v_{f,i}$ and $\hat{v}_{f,i}$ are the reference and predicted values for entry i in the aligned pair, respectively.

Within One Order-of-Magnitude ($p1\mathrm{mag}$) is the fraction of available predictions in column f that lie within one order of magnitude of the reference value:

$$p1\mathrm{mag}_f = \frac{100}{|A_f^+|} \sum_{i \in A_f^+} \mathbf{1}\Big(\big|\log_{10} \hat{v}_{f,i} - \log_{10} v_{f,i}\big| \leq 1\Big), \tag{9}$$

where $A_f^+ = \{i \in A_f \mid v_{f,i} > 0 \text{ and } \hat{v}_{f,i} > 0\}$ is the subset of available predictions with positive values required for the logarithm, and entries with zero or negative values are excluded from this metric only.

C Detailed Experimental Settings and Results

C.1 Experimental Settings

We structure outputs in JSON and guide decoding using LangChain's `create_structured_output_runnable` with a Pydantic schema. This approach enforces consistent formatting and reduces errors (Geng et al., 2025).

We conduct the experiments using a zero-shot condition and generalise the simple prompt for use across all domain-specific datasets in our study. We set the LangChain chunk size to 2000 tokens with a 1% overlap ratio following Ghosh et al. (2024). To ensure stable and reproducible output, we set the temperature parameter to 0.

The prompt used in table generation:

```
You are a helpful assistant capable of extracting information from text. Only extract relevant information from the text. Extract nothing if no important information can be found in the text.
Input: {{Paper (text)}}
Schema: {{JSON schema}}
```

For vision-based generation, we follow the prompt used in the model.

Prompt:

```
Extract the text from the above document as if you were reading it naturally. Return the tables in html format. Return the equations in LaTeX representation. If there is an image in the document and an image caption is not present, add a small description of the image inside the <img></img> tag; otherwise, add the image caption inside <img></img>. Watermarks should be wrapped in brackets. Ex: <watermark>OFFICIAL COPY</watermark>. Page numbers should be wrapped in brackets. Ex: <page_number>14</page_number> or <page_number>9/22</page_number>.
Input: {{Image url}}
```

Computing Costs and Resources

In this study, the approximate cost of using o1-mini was $672 USD. Inference with GPT-OSS 120B, LLaMA-3.3-70B and Nanonets was conducted locally on two Nvidia H100 GPUs.

C.2 Different Parsers

We explored alternative PDF parsing approaches using OCR-based tools like Nanonets-OCR-s (Mandal et al., 2025), Mathpix and PyMuPDF, as some have reported success with these tools (Jiang et al., 2025). Despite known issues, GROBID remains the most reliable option for the benchmark.

To investigate the impact of parsers on table generation performance, we experiment with Docling (Pfitzmann et al., 2022; Team, 2024; Nassar et al., 2022)[21] and Nanonets-OCR-s (Mandal et al., 2025)[22] , the latter being an OCR-based model, in addition to GROBID. Since we cannot do intrinsic evaluation by comparing the extracted text from parsers with the ground truth text, we use the STG task as an extrinsic evaluation setting, where PDF parsing quality directly affects downstream performance. The experiments are conducted using LLaMA 3.3 as the generation model. The results, shown in Table 4, indicate that GROBID achieves the strongest overall performance, although the gap between parsers is relatively small.

[21] https://huggingface.co/ds4sd/docling-models

[22] https://huggingface.co/nanonets/Nanonets-OCR-s

Domain	Dataset	GROBID				nanonets/Nanonets-OCR-s				ds4sd/docling-models			
		ChrF	Precision	Recall	F_1	ChrF	Precision	Recall	F_1	ChrF	Precision	Recall	F_1
Machine Learning	TDMS	11.46	1.99	18.38	**3.59**	11.49	1.11	19.39	2.09	11.30	1.20	20.03	2.26
	SciREX	6.95	1.48	8.93	**2.54**	6.13	1.00	10.43	1.83	6.34	0.89	9.70	1.63
	LEGOBENCH	11.10	2.42	13.18	**4.10**	10.45	2.32	12.56	3.92	10.86	2.34	14.12	4.01
Material Science	MPEA	2.00	10.20	37.83	**16.07**	1.85	6.90	39.15	11.74	1.90	7.45	38.87	12.50
	Diffusion	0.88	10.52	10.16	**10.34**	0.90	7.66	13.84	9.86	0.83	8.19	13.43	10.18
	YSHEAY	14.25	1.82	34.37	**3.47**	14.85	1.00	31.50	1.95	15.22	1.03	30.63	1.99
	CCRMG	10.49	0.70	18.56	**1.35**	11.10	0.68	17.16	1.31	11.32	0.61	16.18	1.18
	Doping	12.52	13.38	20.83	**16.29**	12.58	9.99	22.61	13.86	12.21	11.56	23.23	15.44
	MMD	1.09	8.77	3.57	**5.08**	0.89	4.94	3.04	3.76	0.87	5.75	2.54	3.53
	MRL	0.72	2.44	2.82	**2.62**	0.67	2.00	3.50	2.54	0.70	2.21	3.02	2.55
	PNCExtract	9.40	11.73	27.96	**16.53**	9.37	9.68	34.49	15.12	9.19	9.74	32.20	14.96
	PolyIE	19.02	11.74	27.00	**16.36**	18.95	9.94	28.30	14.71	18.22	10.16	27.39	14.82
Chemistry	BRENDA-enzyme	2.53	11.07	30.72	16.28	2.67	11.81	33.09	**17.41**	2.55	11.90	31.97	17.34
	BRENDA-ribozyme	1.37	4.72	11.39	**6.68**	1.45	4.04	12.51	6.11	1.31	4.25	11.36	6.19
	OPE	4.76	4.66	37.70	**8.29**	4.43	3.29	34.79	6.02	4.68	3.90	38.38	7.08
	PPE	10.75	8.86	61.32	**15.48**	10.83	6.84	60.57	12.29	10.81	7.29	60.94	13.02
	SE	9.74	18.40	30.41	**22.93**	9.39	14.91	33.80	20.69	9.56	15.27	35.40	21.34
Medicine	AE	6.49	7.34	6.84	7.08	10.19	11.44	20.13	**14.59**	7.41	6.02	9.11	7.25
Agriculture	FE	2.50	1.49	28.20	**2.83**	2.31	1.06	29.10	2.04	2.23	0.98	27.69	1.89
Physics	SuperMat	9.04	12.49	19.23	**15.15**	8.68	9.11	15.07	11.36	9.24	3.06	5.06	3.81
	Avg.	7.35	7.31	22.47	**9.65**	7.46	5.99	23.75	8.66	7.34	5.69	22.56	8.15

Table 4: Performance comparison of parsers using LLaMA 3.3. All scores are reported on a 0–100 scale, with the best F_1 score for each dataset highlighted in **boldfaced**.

C.3 Domain Level Results

Figure 5 shows radar plots of model performance across six scientific domains, with metrics averaged at the domain level. The results reveal a consistent pattern: all models achieve higher recall but lower precision, showing that while many candidate records are generated, their accuracy remains limited. Recall is also influenced by the density of records in each dataset. Datasets with fewer records per document tend to have higher recall, as models generate more candidate records.

C.4 Quantity-based results

Quantity-based results are presented in Table 5.

D Explorations on STG Open Challenges

D.1 Entity Normalisation

Entity normalisation is the task of mapping semantically equivalent entity mentions in the generated table to a canonical form. In our setting, mentions appear at the *cell level* of the generated tables. Formally, given a set of cell-level entity mentions $E = e_1, e_2, \cdots, e_n$, the goal is to normalise each e_i to a unique canonical entity $c_j \in C$, where C is the set of standard forms.

Our benchmark datasets include normalised entities to support this task (e.g, SciREX, TDMS, FE). Unsupervised methods such as string similarity metrics (e.g., Levenshtein distance or token-based ratios), string matching libraries (e.g., `TheFuzz`[23]), semantic similarity methods (e.g., embedding-based approaches), or entity ranking algorithms (Sung et al., 2020) can provide effective baseline solutions. These approaches rely on different types of similarity (surface-form or semantic) and can be useful depending on how the mention and the canonical entity vary (e.g., "TiO_2" vs. "Titanium dioxide" or "polymer chain" vs. "macromolecule").

Method: We apply a semantic similarity approach for entity normalisation. We use `PapersWithCode`[24], a comprehensive resource on leaderboard entities, to get normalised forms of entities in the TDMS dataset, which itself adopts PapersWithCode for normalisation. Embeddings are computed for both the extracted values and the candidate normalised entities using the pretrained model[25]. Cosine similarity is then applied to map each extracted value to its closest normalised form. This normalisation post-processing results in a substantial performance gain, with the F_1 score increasing from 3.59 to 18.56.

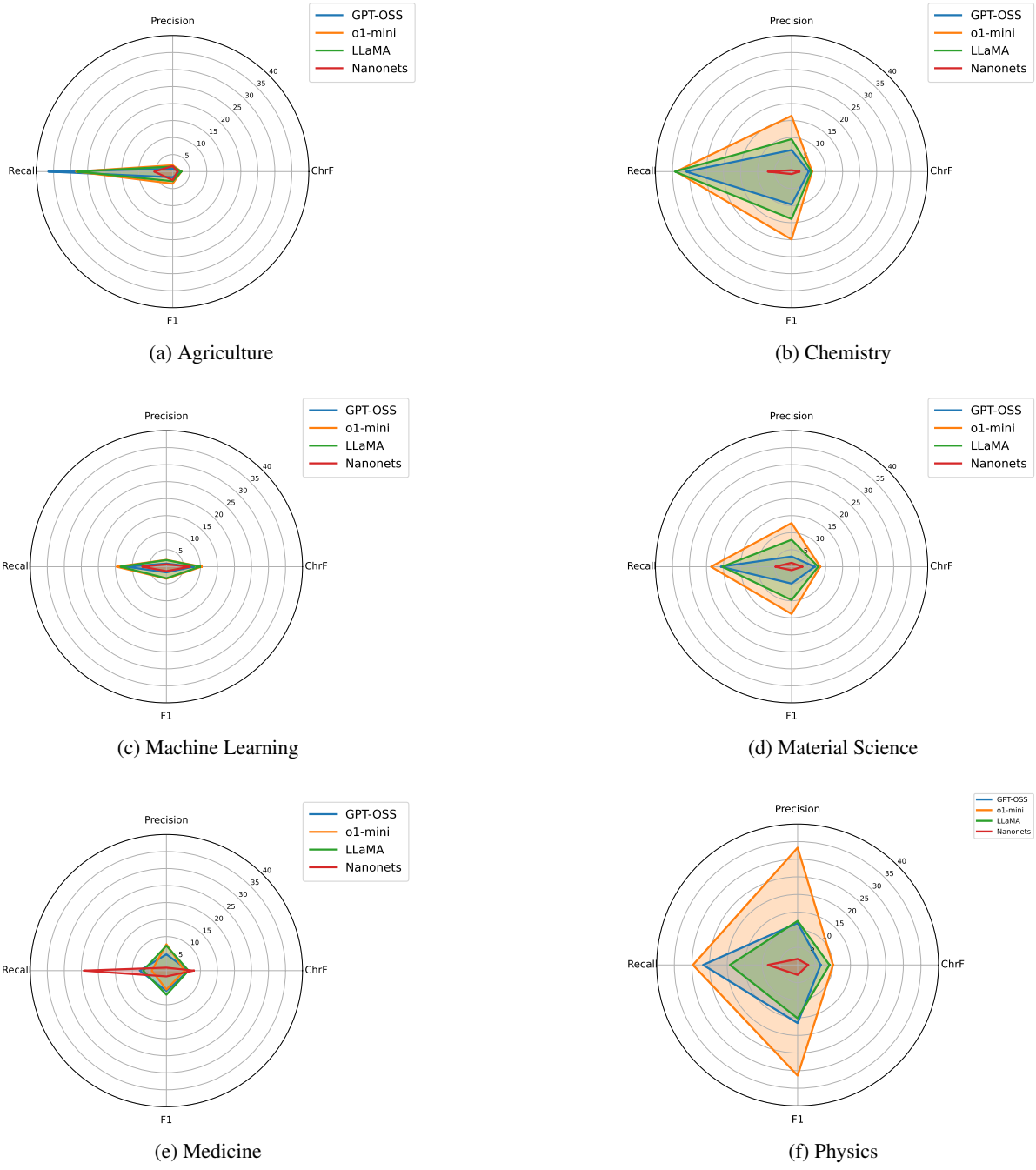
D.2 Unit and Quantity Standardisation

Unit & quantity standardisation is the task of converting quantitative values into a common unit and

[23] https://github.com/seatgeek/thefuzz

[24] https://github.com/paperswithcode/paperswithcode-data

[25] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2



(a) Agriculture (b) Chemistry

(c) Machine Learning (d) Material Science

(e) Medicine (f) Physics

Figure 5: Radar plots showing the performance of GPT-OSS, o1-mini, LLaMA 3.3, and Nanonets (vision) across domains. Metrics include ChrF, Precision, Recall, and F_1, averaged over all datasets in each domain.

standardised numeric format. Given a set of quantity pairs $Q = (v_1, u_1), \cdots, (v_n, u_n)$ extracted from document d, where v_i is a numeric value and u_i is its associated unit (stored in separate columns in the generated table), the goal is to convert each (v_i, u_i) into a standardised pair $(v_i^{'}, u_i^{'})$ using unit conversion rules.

Our benchmark datasets provide raw-value-unit attributes to support this task (BRENDA-enzyme and BRENDA-ribozyme). Simple methods include rule-based conversion systems (Jiang et al., 2025), curated unit mappings, or libraries like `pint`[26].

Method: We apply unit normalisation using `pint` library on the BRENDA-enzyme dataset, where units are annotated as a separate attribute (see Table 3). In this dataset, K_m values are standardised to millimolar (mM) and k_{cat} values to per minute (min^{-1}). The `pint` library enables convert-

[26] https://github.com/hgrecco/pint

Domain	Dataset	Attribute	Coverage (↑)	MAE (↓)	$p\mathbf{1}mag$ (↑)
Machine Learning	TDMS	Score	65.76	30.32	81.56
	SciREX	Score	63.30	496.51	78.82
	LEGOBENCH	Score	70.10	703213.97	76.71
Material Science	MPEA	grain size	40.89	52.14	78.26
		experimental density	11.68	571.10	88.54
		test temperature	37.46	453.88	53.48
		yield strength	67.37	241.70	94.39
		ultimate tensile strength	54.62	237.17	93.31
		elongation	37.92	12.94	81.86
		elongation plastic	40.43	2.49	87.72
	Diffusion	test temperature	75.94	331.87	98.32
		pressure	50.94	1332.31	28.35
		diffusivity	43.20	11.72	17.09
		SiO2	21.45	9.75	90.15
		TiO2	18.83	0.39	84.14
		Al2O3	16.39	2.40	97.51
		FeOt	18.58	2.24	84.42
		MnO	20.25	0.01	99.24
		MgO	17.75	1.07	95.77
		CaO	17.07	2.04	95.41
		Na2O	16.59	0.98	98.64
		K2O	16.02	0.31	99.46
		P2O5	15.62	0.01	100.00
		H2Ot	36.02	1.40	85.96
	YSHEAY	Value	91.36	246.60	86.11
	CCRMG	Value	71.08	515.17	74.58
	Doping	—	—	—	—
	MMD	Formula units per cell	26.92	1.57	100.00
		Atomic sites per cell	11.54	2.67	100.00
		Formation energy (eV/atom)	23.08	16.41	0.00
		Energy relative to convex hull (eV/atom)	30.77	20.35	0.00
		Averaged magnetic moment (μB/atom)	30.77	0.72	87.50
		Magnetic polarization, Js</(T)	11.54	2.67	66.67
		Magnetic anisotropy constants:Kb-c (MJ/m3)	50.00	1.14	50.00
	MRL	Electrical resistivity (Ωcm)	20.98	38.09	7.81
		Seebeck coefficient (μCV/K)	29.51	147.99	72.22
		ZT (thermoelectric figure of merit)	46.56	0.63	83.80
		temperature	71.15	272.25	95.39
		Thermal conductivity (W/mK)	34.38	13.80	83.12
		Average atomic mass (g/mol)	0.33	0.24	100.00
		Atoms per unit cell	3.28	4.73	100.00
	PNCExtract	Filler Composition Mass	81.06	7.88	9.03
		Filler Composition Volume	77.10	6.43	19.39
	PolyIE	Property Value	91.73	43.91	76.13
Chemistry	BRENDA-enzyme	km	57.83	61763.09	43.92
		kcat	56.21	4814.92	71.45
		km_kcat	53.85	633623.38	35.93
		Commentary[Temp]	30.85	4.12	99.49
		Commentary[pH]	26.16	0.41	99.87
	BRENDA-ribozyme	km	57.98	183.64	86.47
		kcat	58.94	5.61	83.61
		km_kcat	29.62	122445.26	74.31
		Commentary[Temp]	36.21	4.50	99.29
		Commentary[pH]	44.39	0.38	100.00
	OPE	Von [V]	66.87	1660.93	76.58
vim	PPE	PCE_max(%)	89.81	0.28	100.00
		PCE_ave(%)	69.62	0.28	100.00
		Voc (V)	79.25	0.06	99.52
	SE	temperature	80.81	34.12	99.02
		pressure	74.03	160.54	94.42
		solubility	64.18	643.14	77.95
Medicine	AE	Affinities	72.83	33849.81	54.48
Agriculture	FE	screw dimension	23.91	40.12	90.91
Physics	SuperMat	tcValue	66.74	9.14	92.64
		Pressure	82.08	2.99	89.66

Table 5: Quantitative evaluation of numerical attributes using o1-mini predictions. **Coverage** indicates the proportion of numerical values predicted, **MAE** denotes Mean Absolute Error, and $p\mathbf{1}mag$ denotes Within One Order-of-Magnitude. Note that the Doping dataset has no numerical attributes.

ing units (e.g., K_m: 15 $\mu M \rightarrow 0.015\ mM$). Evaluation on these quantitative attributes (K_m and k_{cat}) demonstrates an improvement in performance, with the F_1 score increasing from 34.2 to 38.7 after unit normalisation.

D.3 Entity Linking

Entity linking is the task of mapping entity mentions in the generated table to entries in an external knowledge base. Given a set of mentions $M = m_1, m_2, \cdots, m_n$ and a knowledge base $\mathcal{K}$, the goal is to identify a mapping function $f : M \rightarrow \mathcal{K}$ such that $f(m_i)$ returns the correct canonical entity from $\mathcal{K}$ for each mention m_i.

Our benchmark includes machine learning datasets (TDMS, SciREX and LEGOBENCH), which contain *task*, *dataset*, and *metric* attributes presented in their canonical forms and can be evaluated for this subtask. One of the solutions for entity linking includes computing semantic similarity between mentions and candidate entries using pretrained embedding models (Karadeniz and Özgür, 2019; Yamada et al., 2020).

Method: We apply the same normalisation method for entity linking on the TDMS dataset.

D.4 Table and Figure Understanding

Table and figure understanding is the task of extracting structured information from tables and figures in scientific documents. Given a visual input x (a table or figure) embedded in a document d_i, the goal is to parse its structure and extract content to enrich the textual representation of d_i. For tables, this typically involves recovering explicit information such as numerical values, labels, and table structures. For figures, the extracted information may be explicit (e.g., axis labels, data points) or implicit (e.g., observed trends, comparisons, or relationships). Understanding tables and figures is also important for tasks such as reasoning (Lu et al., 2025), text generation (Moosavi et al., 2021; Chen et al., 2021), interpretation (Roberts et al., 2024), and question answering (Borisova et al., 2025), as it enables models to integrate visual and textual information for comprehensive scientific analysis.

All datasets in our benchmark are suitable for the table understanding subtask. For the figure understanding task, material science datasets (e.g., Diffusion, MPEA, and MRL) can be used since they require extensive figure understanding for the STG task. However, our benchmark does not provide ground-truth annotations for table or figure understanding. Therefore, evaluation of these subtasks can only be performed extrinsically, via their impact on STG.

D.5 Quantitative Value Evaluation

Quantitative value evaluation is the task of assessing the correctness of extracted numerical values, accounting for unit normalisation, numeric format, and acceptable tolerances. Given a predicted value $\hat{v}$ and reference value v, the goal is to determine if $\hat{v}$ matches v within a defined equivalence function $f_{\text{eq}}(\hat{v}, v) \in 0, 1$ that incorporates value equivalence, unit consistency, and numeric precision.

Method: We evaluate numerical correctness using quantity-based metrics defined in the evaluation framework (see Section 5), and the results are presented in Table 5.