METALEAD: A Comprehensive Human-Curated Leaderboard Dataset for Transparent Reporting of Machine Learning Experiments

Roelien C. Timmer **Necva Bölücü** **Stephen Wan**
CSIRO Data61, Australia
{roelien.timmer, necva.bolucu, stephen.wan}@data61.csiro.au

Abstract

Leaderboards are crucial in the machine learning (ML) domain for benchmarking and tracking progress. However, creating leaderboards traditionally demands significant manual effort. In recent years, efforts have been made to automate leaderboard generation, but existing datasets for this purpose are limited by capturing only the best results from each paper and limited metadata. We present METALEAD, a fully human-annotated **ML leaderboard** dataset that captures all experimental results for *result transparency* and contains extra **metadata**, such as the result *experimental type*(baseline, proposed method, or variation of proposed method) for *experiment-type guided comparisons*, and explicitly separates *train* and *test dataset* for *cross-domain assessment*. This enriched structure makes METALEAD a powerful resource for more transparent and nuanced evaluations across ML research. METALEAD dataset and code repository: https://github.com/RoelTim/metalead

1 Introduction

The volume of scientific literature has grown exponentially[1], resulting in increased *cognitive load* for researchers, who struggle to compare existing work due to a *scientific information overload* (Landhuis, 2016; Bornmann et al., 2021). This information overload for researchers, combined with the fast-paced development of AI, has led to a growing supply of *AI for Science* tools designed to help researchers efficiently process and understand existing research, such as *Google's Co Scientist* (Gottweis et al., 2025), *Ought's Elicit* (Elicit, 2023), and, *OpenAI's Deep Research* (OpenAI, 2025).

For data science workflows, we note that the field already employs various conventions to assist in research planning and evaluation. For example, in Machine Learning (ML), a tool to easily compare the performance of different models is *ML leaderboards*. A popular online platform for ML leaderboards is *Papers with Code*[2] , which links research papers to their associated code implementations and maintains community-curated leaderboards across tasks..

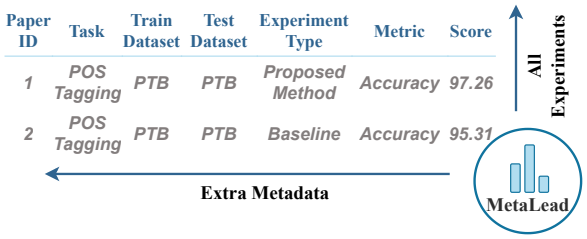

Figure 1: METALEAD is more comprehensive than existing leaderboard datasets, with broader metadata (horizontal) and full inclusion of experimental results (vertical).

However, manually curating these types of leaderboards is labour-intensive and requires continual manual maintenance. To ease this burden and enable *cognitive offloading*[3] (Risko and Gilbert, 2016), researchers have developed NLP-based tools that automate the curation process (Singh et al., 2019; Hou et al., 2019; Kardas et al., 2020; Kabongo et al., 2021; Yang et al., 2022; Kabongo et al., 2024a). With the first publication in 2019, based on a BERT-style transformer encoder trained for natural language inference (NLI) (Hou et al., 2019), while recent publications are based on multi-agent LLM frameworks (Şahinüç et al., 2024).

Even though NLP methods for extracting ML leaderboard entries have advanced, the scope of leaderboards remains limited (Timmer et al., 2025). Existing leaderboards (§ 2) report only the best results, omitting ablations, negative results, and base-

[1]The scale of this increase is well demonstrated by arXiv, with a consistent annual growth in publication volume of approximately 10% (arXiv, 2025).

[2]https://paperswithcode.com/

[3]Cognitive offloading refers to the use of external tools or strategies to reduce mental effort.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4190–4206

Dataset	All Results	Annotation Source	Scope Tuples
Hou et al. (2019)	✗	NLP Progress	⟨*task, dataset, metric, score*⟩
Kardas et al. (2020)	✗	Papers with Code	⟨*task, dataset, metric, score*⟩
Kabongo et al. (2024a)	✗	Papers with Code	⟨*task, dataset, metric, score*⟩
Singh et al. (2024)	✗	Papers with Code	⟨*task, dataset, metric, method, score*⟩
Şahinüç et al. (2024)	✗	Human Annotation	⟨*task, dataset, metric, score*⟩
METALEAD (Ours)	✓	Human Annotation	⟨*task, train dataset, test dataset, metric, score, experiment type*⟩

Table 1: Comparison of METALEAD and related datasets. METALEAD uniquely includes all reported results, uses human annotation, and captures extended tuples with dataset splits and *experiment type*.

lines. For instance, a recent shared task extracted only ⟨*task, dataset, metric, score*⟩ tuples (D'Souza et al., 2024), reflecting the field's current focus while neglecting broader experimental context.

The limited scope of current leaderboards leads to two key *shortcomings*. First, focusing only on best-performing results causes *performance myopia*[4] (Ethayarajh and Jurafsky, 2020) and promotes publication bias, favouring positive results (Karl et al., 2024). Second, the lack of distinction between train and test datasets hinders meaningful *in-domain* (ID) versus *out-of-domain* (OOD) comparisons, a critical distinction, as performance often varies across domains, leading to different conclusions and downstream consequences (Teney et al., 2020; Gebru et al., 2021).

To address these shortcomings, we introduce METALEAD, building on the previously annotated document set *SciLead* introduced by Şahinüç et al. (2024), a manually annotated benchmark of scientific leaderboards that captures (task, dataset, metric, score) tuples from ML papers. Our additional annotations are illustrated in Figure 1, adding 3,285 new records to the original 295 best records (i.e., leaderboard result) of Şahinüç et al. (2024). Although METALEAD diverges from conventional leaderboards by including all reported results, it preserves the leaderboard's core purpose: enabling comparative evaluation of model performance. This broader inclusion enhances transparency rather than undermining the concept.

The three main contributions of this work are as follows: First, we introduce **METALEAD**, the first leaderboard-style dataset to include *all* experimental results from papers, enabling more transparency (§ 3.1). Second, we expand the schema of leaderboard tuples in two ways: (1) by introducing an explicit *experiment type* field, which enables experiment-type guided comparisons (§ 3.2) and is accompanied by a novel metric for evaluation (§ 5.3); and (2) by adding an explicit *train/test dataset distinction*, which supports cross-domain assessment (§ 3.3). Third, we provide a fully **human-annotated** resource (§ 4) and benchmarking results across various LLMs (§ 5-§ 6). Overall, METALEAD represents a step towards more holistic ML evaluation, broadening the notion of leaderboards and enabling more rigorous and transparent scientific inquiry.

2 Related Work

ML Leaderboard Datasets METALEAD builds on prior work in ML leaderboard generation. Table 1 summarises existing datasets and shows how METALEAD compares. The comparison includes only datasets that report experimental results, as leaderboards are only meaningful when scores are provided.[5] A key distinction of METALEAD is its comprehensive coverage: METALEAD includes all experimental results, whereas prior datasets report only the best results. Most existing datasets rely on community annotations, for example Papers with Code and its precursor NLP Progress[6] In contrast, only METALEAD and Şahinüç et al. (2024) provide fully human-annotated leaderboards, ensuring high-quality annotations. Tuple scope further differentiates datasets: most extract ⟨*task, dataset, metric, score*⟩ tuples, while Singh et al. (2024) reports ⟨*method, score*⟩ conditioned on ⟨*task, dataset, metric*⟩. METALEAD's structure is the most detailed: ⟨*task, train dataset, test dataset, metric, score, experiment type*⟩, distinguishing training and test datasets and categorising results as baselines, proposed methods, or variations.

Scientific Information Extraction Automatic ML leaderboard generation is part of a wider body

[4]Performance myopia is the tendency to focus solely on best-performing results, neglecting broader context such as negative results, baselines, or failed variants.

[5]E.g., Kabongo et al. (2021) report ⟨*task, dataset, metric*⟩ tuples but do not include scores.

[6]https://nlpprogress.com/, a GitHub community-maintained NLP leaderboard.

of work about scientific information extraction (IE). Jain et al. (2020) introduced SciREX, a benchmark dataset for scientific relation extraction focusing on identifying experimental results and their context within scientific documents. Similarly, Hou et al. (2021) proposed TDM-Sci, which focuses on structured extraction of ⟨*task, dataset, metric*⟩ triples from scientific papers. Other examples are IE used for scientific knowledge graph construction (Ammar et al., 2018; Luan et al., 2018; Mondal et al., 2021; Timmer et al., 2023; Khoo et al., 2023) and the extraction of citations and entities (Jurgens et al., 2018)

3 Schema Design Choices for METALEAD

METALEAD is designed to extend the leaderboard format to serve as a *research planning scaffold* for interpreting experimental results. METALEAD builds on existing human-curated leaderboard datasets by promoting *result transparency* (§ 3.1), *experiment-type guided comparison* (§ 3.2), and *cross-domain assessment* (§ 3.3). Each design choice addresses specific bottlenecks in interpreting ML experiments at scale, which have been previously identified and discussed in the literature. Figure 2 shows a diagram that visualises how these three design choices impact leaderboards.

3.1 Result Transparency

All existing leaderboard datasets include only the best-performing results, promoting *performance myopia* and *reinforcing confirmation bias* (Ethayarajh and Jurafsky, 2020). Science suffers from publication bias, favouring positive results, and leaderboards exacerbate this by showcasing only top results (Karl et al., 2024). Previous ML generation systematically point out that a key direction for future research is including all experimental results, for example Timmer et al. (2025) advocate for vertical expansion of leaderboards, Hou et al. (2019) call it an ideal situation to extract all results reported in the same paper, Kardas et al. (2020) point out that retrieving all results rather than just the principal results introduced in the paper would enable fine-grained comparisons.

Reporting bias is a big problem in the ML field, for example McGreivy and Hakim (2024) found that 79% of papers compare only against weak baselines. Including all baselines in leaderboards, such as in METALEAD, increases transparency and reduces the risk that researchers inadvertently or deliberately select weak baselines.

In contrast, the METALEAD schema captures *all* reported results, including ablations, negative outcomes, and failed variations. This component supports what we term *result transparency*: the idea that decision quality improves when users are exposed to all the experimental results rather than a curated success narrative.

3.2 Experiment-Type Guided Comparisons

Including all experimental results can overwhelm users if not paired with an appropriate experimental description. Previous work vow for enriching leaderboard by providing more experimental type information, for example, Kabongo et al. (2024b) are interested in extend the current triples (task, dataset, metric) model with additional concepts, Timmer et al. (2025) recommend expanding leaderboards horizontally, and Kardas et al. (2020) mentions future work may want to build on our approach for more comprehensive extraction task. However, none of these works is detailed about what exactly this expansion should entail. To facilitate meaningful comparisons, the METALEAD schema classifies each result by its *experiment type*: *baseline*, *proposed method*, or *variation of proposed method*. This categorical distinction provides *experimental context* that helps researchers organise the landscape of results and interpret empirical evidence within its proper context. By explicitly encoding the role of each result, the schema reduces ambiguity, supports faster categorisation and enables faster insight.

3.3 Cross-Domain Assessment

Existing leaderboard datasets' schemas force the extraction of a single dataset mention, introducing an ambiguity regarding whether the mention pertains to the training and evaluation contexts. This obscures whether the evaluation results are In-Domain (ID) or Out-of-Domain (OOD). Prior research has shown that ID and OOD evaluations can yield markedly different results (Teney et al., 2020; Gebru et al., 2021). Liu et al. (2021) of ExplainaBoard, advocates a partitioning of the test set, allowing researchers to observe where models perform well or poorly and to gain insights beyond overall accuracy scores. The METALEAD schema addresses this by explicitly separating the *train* and *test* datasets. By listing the train and test sets separately in METALEAD, researchers can more easily identify patterns, e.g., the agreement-on-the-line

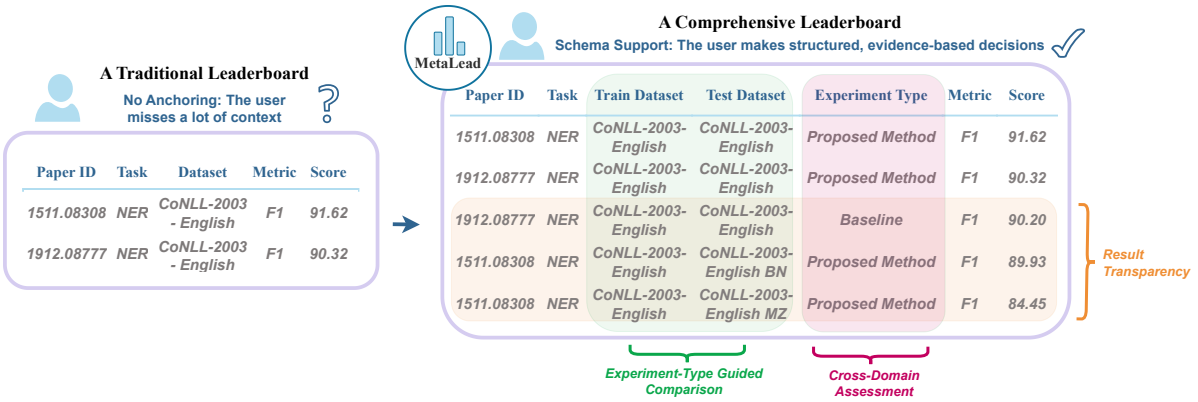

Figure 2: Schema benefits introduced by METALEAD. Compared to traditional leaderboards, METALEAD enables structured analysis through three key schema innovations: 1) *result transparency* by including all reported outcomes, 2) *experiment-type guided comparisons* via explicit labelling of baselines, proposed methods, and variations, and 3) *cross-domain assessment* by distinguishing between training and test datasets.

	Count	
Dataset Item	**SciLead**	**METALEAD**
General Statistics		
Papers	43	43
Total tuples	295	3,568
Avg. Annotations per Paper	7	83
Median Annotations per Paper	6	61
Diversity of Content		
Unique Leaderboards	27	283
Unique Tasks	23	23
Unique Train Datasets	-	83
Unique (Test) Datasets	71	160
Unique Metrics	26	26
Experimental Type		
Baseline	-	1,919
Proposed Method	-	809
Variation of Proposed Method	-	840

Table 2: Statistics of our METALEAD dataset versus SciLead. **-** if not defined, as these are not existing categories/labels in the SciLead dataset.

described by Baek et al. (2022), a distributional robustness phenomenon regarding ID and OOD in ML, which we describe as *cross-domain assessment*.

4 METALEAD Dataset Annotation

Table 2 presents the statistics of the fully human-annotated METALEAD dataset, constructed through a two-step annotation process: step 1: core tuple annotation (§ 4.1) and step 2: experiment type labelling (§ 4.2). We outline key annotation challenges encountered during the process (§ 4.3), and describe how the leaderboards are constructed from the annotated tuples (§ 4.4).

4.1 Core Tuple Annotation

The annotation process was carried out in two distinct steps. The annotators are the authors of this paper, further details are provided in the Ethics section. Step 1 is annotating the core tuple: ⟨*task, train dataset, test dataset, metric, score*⟩. In line with the quality-control single-annotator approach introduced by Şahinüç et al. (2024), the annotator annotates all the experimental results reported in the paper. The annotation instructions for step 1 are provided in Appendix A Figure 4.

4.2 Experiment Type Labelling

Step 2 is labelling the *experiment type*. Figure 3 shows the decision tree for annotating *experiment type* and the detailed annotation instructions are provided in Appendix A Figure 5. To ensure clarity and consistency for this *potentially* more subjective classification task, two annotators independently annotated an initial batch of five randomly selected papers, we observed a Cohen's Kappa of 0.91, indicating *near-perfect agreement* (Cohen, 1960). We then held a resolution discussion to address the disagreements, during which we revised both the

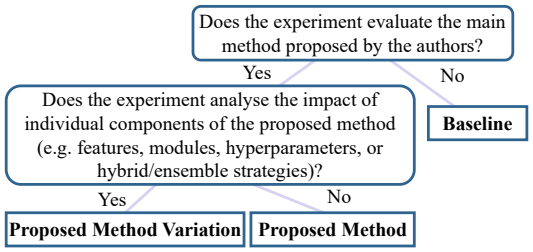

Figure 3: Decision tree for the *experiment type* entry of METALEAD.

annotation guidelines and the corresponding annotations, resulting in perfect agreement (Cohen's Kappa = 1.00). Using the updated guidelines, we annotated a second batch of five papers and again observed near-perfect agreement (Cohen's Kappa = 0.94). Given the consistently high inter-annotator agreement, the remaining papers were annotated by a single annotator using the finalised guidelines.

4.3 Annotation Challenges

Annotation of papers for leaderboards presents several challenges. A common issue is that authors are not explicit about whether they propose a single main method or multiple methods. For example, Narayan et al. (2018) propose four variants of a *T-CONVS2S* architecture, but the authors do not make clear whether one variant is the main method or whether all should be considered equivalent. An example of inconsistency within a paper appears in the work by Li et al. (2019), where the authors describe a method as a baseline in the main text, yet label it as "Ours" rather than "Baseline" in the corresponding results table. These inconsistencies complicate the annotation process and introduce ambiguity in classifying experimental results.

4.4 Leaderboard Construction

We construct a leaderboard for each unique combination of ⟨*task, train dataset, test dataset, metric, score*⟩, resulting in a total of 283 leaderboards. Whereas some existing leaderboard datasets impose a minimum threshold to consider a leaderboard valid (e.g., Kabongo et al. (2021) require five papers per leaderboard), we do not apply such thresholds. The aim of METALEAD is to offer a more granular view of experimental results. Imposing a threshold would have excluded important information, particularly for settings such as ID versus OOD evaluation.

5 Evaluation

We evaluate extraction performance in open and closed domains (§ 5.1) using tuple- and entity-level metrics (§ 5.2), propose a new metric for *experiment type* classification accuracy (§ 5.3), and introduce leaderboard-specific metrics to assess broader utility (§ 5.4).

5.1 Open and Closed Domain Scenario

Following Timmer et al. (2025), we evaluate both open and closed domain settings.[7] In the open domain, the model receives no candidate tuples and must extract all relevant entries directly from the paper. In the closed domain, the model selects entities from a predefined list. For open domain evaluation, we apply paper-level normalisation, as described by Şahinüç et al. (2024).

5.2 Extraction Level Metrics

Given the novel nature of a task for comprehensive extraction of results, we introduce and define our evaluation metrics here. We report the *set-based* recall, precision, and F1, *averaged across all papers*. Recall (**R**) is given by:

$$\mathbf{R} = \frac{\sum_{p \in P} \frac{|E_p^G \cap E_p^M|}{|E_p^G|}}{P} \quad (1)$$

for all papers P, recall is computed per paper $p \in P$ as the ratio of correctly predicted entities (or tuples), $|E_p^G \cap E_p^M|$, to the total number of gold entities, $|E_p^G|$. Here, $|\cdot|$ denotes the *cardinality* of a set, i.e., the number of unique elements. The overall recall is the average recall across all papers.

Precision (**P**) is given by:

$$\mathbf{P} = \frac{\sum_{p \in P} \frac{|E_p^G \cap E_p^M|}{|E_p^M|}}{P} \quad (2)$$

with for each paper $p \in P$, precision is computed as the ratio of correctly predicted entities $|E_p^G \cap E_p^M|$ to the total number of predicted entities $|E_p^M|$.

Based on **R** and **P**, we also report the **F1** score. Together, the set-based **R**, **P**, and **F1** metrics assess how well the model captures relevant entities across the corpus, while avoiding overcounting due to repeated predictions. They offer a more conservative measure of performance by focusing on the presence of correct entities rather than frequency.

5.3 Experiment Type Evaluation

We propose a novel metric to evaluate *experiment type* classification: for each correctly extracted ⟨*task, train dataset, test dataset, metric, score*⟩ tuple, we compute the percentage with correctly assigned *experiment type*, **ET-Accuracy**:

$$\textbf{ET-Accuracy} = \frac{\sum_{p \in P} \frac{|\hat{T}_p \cap T_p|}{|\hat{T}_p|}}{P} \quad (3)$$

[7] Also known as the *cold-start* scenario (Şahinüç et al., 2024).

where $\hat{T}_p$ is the set of predicted tuples for paper p with a correct core tuple, and $T_p \subseteq \hat{T}_p$ contains those with a correct *experiment type* label. To better understand classification difficulty, we compute confusion matrices. Each confusion matrix is normalised per true label (i.e., each row sums to 100%), and we report the average across all non-empty matrices, where rows with no true instances are excluded from the average.

5.4 Leaderboard Level Metrics

Following Şahinüç et al. (2024), we calculate four leaderboard-specific metrics: *Leaderboard Recall*, *Paper Coverage*, *Result Coverage*, and, *Average Overlap* .

We define $\mathcal{L}$ as the set of gold leaderboards. For each leaderboard $\ell \in \mathcal{L}$, let P_ℓ^{gold} and P_ℓ^{pred} be the sets of gold and predicted papers. Leaderboard Recall (**LR**) is calculated by seeing how many of the leaderboards have been created, *i.e.,* how many of the unique golden ⟨*task, train dataset, test dataset, metric, experiment type*⟩ tuples have been reported in the predicted:

$$\textbf{LR} = \frac{|\mathcal{L}^{\text{gold}} \cap \mathcal{L}^{\text{pred}}|}{|\mathcal{L}^{\text{gold}}|} \tag{4}$$

Paper Coverage (**PC**) is the average recall of assigned papers[8]:

$$\textbf{PC} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \frac{|P_\ell^{\text{gold}} \cap P_\ell^{\text{pred}}|}{|P_\ell^{\text{gold}}|} \tag{5}$$

Result Coverage (**RC**) is computed analogously, where R_ℓ^{gold} denotes the set of ground truth results on leaderboard ℓ, and R_ℓ^{pred} denotes the set of predicted results on leaderboard ℓ:

$$\textbf{RC} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \frac{|R_\ell^{\text{gold}} \cap R_\ell^{\text{pred}}|}{|R_\ell^{\text{gold}}|} \tag{6}$$

Average Overlap (**AO**) (Webber et al., 2010) is a metric to measure the similarity between two ranked leaderboards:

$$\textbf{AO} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \frac{1}{d} \sum_{i=1}^{d} \frac{|r_{\ell,1:i}^{\text{gold}} \cap r_{\ell,1:i}^{\text{pred}}|}{i} \tag{7}$$

where, $r_{\ell,1:i}^{\text{gold}}$ and $r_{\ell,1:i}^{\text{pred}}$ denote the top-i elements of the gold and predicted ranks of leaderboard ℓ.

[8]We treat paper assignments as sets and ignore duplicate entries per leaderboard.

6 Experiments

Experimental Setup We conduct experiments using proprietary, open source[9], and baseline models. Proprietary models include GPT-4.1 (OpenAI, 2024a), GPT-4o (OpenAI, 2024b), o4-mini (OpenAI, 2025), Gemini-2.5-Pro, Gemini-2.5-Flash (Comanici et al., 2025), and Gemini-1.5-Pro (Team et al., 2024). Open source models include Llama 3.3 70B Instruct (Dubey et al., 2024) and Mistral Large (Hugging Face, 2024). Prompts are shown in Appendix B, and more experimental details in Appendix C. We implement modified versions of the baselines by Şahinüç et al. (2024)[10] and Kardas et al. (2020)[11]. For Şahinüç et al., we updated the prompt to match the new schema. For Kardas et al., we extended the code to extract all results, though the output remains limited to ⟨task, dataset, metric, score⟩ tuples. Where possible, we replaced the original LLMs of these baselines with GPT-4.1 for a fairer comparison.

Tuple Extraction Results Table 3 shows exact tuple match results for ⟨*task, train dataset, test dataset, metric, score, experiment type*⟩ in both closed- and open-domain settings. As expected, closed-domain results are higher, though this setting is less realistic. GPT-4.1 and Gemini-2.5-Pro outperform other models. Gemini-1.5-Pro shows high precision but low recall, indicating correct but incomplete extraction. All proprietary models outperform open source models, and all LLMs outperform the baselines (Şahinüç et al., 2024; Kardas et al., 2020). The baselines underperform due to their original design for a different schema and reliance on text filtering strategies, which were necessary for earlier LLMs with smaller context windows. In contrast, current models benefit from full-text processing.

Individual Entities Results Table 4 present the results for the extraction of individual entities in the closed and open domain settings. The performance on individual entities is substantially higher than on complete tuples, as shown in Table 3[12]. This indicates that extracting individual entities is considerably easier than correctly extracting their

[9]We use the term open source, although in practice the openness may be limited, e.g. only weights may be released.

[10]https://github.com/UKPLab/arxiv2024-leaderboard-generation

[11]https://github.com/paperswithcode/axcell

[12]Results for *experiment type* are reported separately in Section Table 6 using a specialised metric.

	Type	Name	P	R	F1
Closed Domain	**Proprietary**	GPT-4.1	46.04	38.16	40.21
		GPT-4o	38.65	9.76	13.93
		o4-mini	45.88	36.79	39.42
		Gemini-2.5-Pro	**46.55**	**52.34**	**48.89**
		Gemini-2.5-Flash	17.78	19.13	18.19
		Gemini-1.5-Pro	42.47	17.42	21.98
	Open Source	Llama 3.3 70B	37.42	24.82	26.96
		Mistral Large	40.43	24.86	28.21
	Baselines	Şahinüç et al. (2024)	18.41	17.78	18.10
		Kardas et al. (2020)*	14.45	79.69	76.37
Open Domain	**Proprietary**	GPT-4.1	16.01	13.49	14.14
		GPT-4o	33.50	9.14	12.61
		o4-mini	37.23	**28.02**	**29.99**
		Gemini-2.5-Pro	24.32	27.02	25.48
		Gemini-2.5-Flash	16.65	17.61	16.81
		Gemini-1.5-Pro	16.54	8.38	9.90
	Open Source	Llama 3.3 70B	**37.51**	23.14	25.78
		Mistral Large	33.77	24.02	26.60
	Baselines	Şahinüç et al. (2024)	1.91	2.14	2.02

* The scope of the tuple for (Kardas et al., 2020) is limited to: ⟨*task, (test) dataset, metric, score*⟩, and therefore these scores are not considered to be top results.

Table 3: Exact Tuple Match for METALEAD for ⟨*task, train dataset, test dataset, metric, score, experiment type*⟩, with top results in the closed domain highlighted in **blue** and in the open domain in **orange**.

combinations. Consistent with the tuple extraction results, GPT-4.1 and Gemini-2.5-Pro achieve the highest scores for individual entity extraction. The lowest scores are generally observed for train and test dataset, likely because the models must distinguish between them. The models also use varying terminology, for example, some models return "CoNLL-2003 English" for both the train and test datasets, while others report "CoNLL-2003 English" for the train dataset and "CoNLL-2003 English (test)" for the test dataset. As expected, the performance for the open domain are much lower, as this is a more challenging and realistic setting.

Leaderboard Results Table 5 presents the results for leaderboard construction. The best-performing model, Gemini-2.5-Pro, achieves a leaderboard recall of only 49.46 in the closed domain, meaning fewer than half of the gold-standard leaderboards are reconstructed. Paper coverage (41.05), result coverage (36.27), and average overlap (7.99) show that even when some leaderboards are reconstructed, the included papers and results only partially match the gold data, and the ranking similarity is very limited. For Kardas et al. (2020), despite some overlap between the gold and predicted leaderboard, it was very limited (0.00). In the more realistic open-domain setting, leaderboard results drop significantly across all models. Both domains show substantial room for improvement.

Experiment Type Results Table 6 reports the correctly classifying on the tuples based on *experiment type*. Şahinüç et al. (2024) achieve the best performance, correctly classifying ∼82% and ∼81% of tuples in the closed and open domains, respectively. One possible explanation for the strong ET-Accuracy of the Şahinüç et al. (2024) baseline is that the method may have extracted only a subset of simpler tuples, where the experimental context was easier to infer. Surprisingly, GPT-4.1 and Gemini-1.5-Pro show a big difference between the closed- and open-domain settings.

Error Analysis Experiment Type To analyse misclassifications in *experiment type*, we create a confusion matrix. The matrix for closed-domain GPT-4.1 is shown in Table 7; others appear in Appendix D. The matrix shows that proposed methods are frequently misclassified as method variations (23.57%), and vice versa (11.55%), while baselines are more reliably identified (89.67% correct). This is likely because authors often use vague phrasing such as "our approach" or "this setting", and do not clearly distinguish between introducing a single method and presenting multiple variations.

7 Discussion

Additional Metadata For future work, we aim to expand the scope of the leaderboard by including additional metadata, such as method names and related attributes. As demonstrated by Otto et al. (2023), this is a challenging task, as their paper shows that labelling ML models the annotator agreement is low (e.g. for *Model Architecture* the annotator agreement was only 23.7 for exact-match and 34.4 for partial-match). Second, terminology and conceptual frameworks may evolve over time, making consistent annotation difficult across papers from different publication periods.

Improved Extraction Framework While individual entity extraction is strong, combining entities correctly remains challenging, as reflected by leaderboard-specific metrics like leaderboard recall. While this paper only investigates zero-shot extraction, future work could explore whether supervised adaptation improves structured result extraction of tuples such as task, dataset, metric,

	Model		Task			Train Dataset			Test Dataset			Metric			Result		
	Type	Name	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Closed Domain	**Proprietary**	GPT-4.1	**97.67**	**99.77**	**98.33**	81.82	78.14	78.86	81.42	71.89	74.02	88.99	82.62	83.93	93.93	67.22	74.16
		GPT-4o	95.40	97.03	95.28	80.84	72.96	75.00	76.58	62.04	65.96	89.26	72.65	77.77	95.23	24.11	34.48
		o4-mini	89.40	97.44	91.71	86.65	81.98	**83.06**	76.72	65.66	68.36	86.07	82.47	82.33	96.82	72.02	77.82
		Gemini-2.5-Pro	90.11	99.51	92.85	81.15	**83.33**	81.58	81.88	**83.42**	**81.59**	78.11	**88.01**	81.32	88.40	**99.07**	**92.74**
		Gemini-2.5-Flash	73.47	85.47	77.40	59.98	68.92	62.33	56.38	58.32	55.83	66.59	73.25	68.18	89.16	95.92	91.28
		Gemini-1.5-Pro	93.02	96.12	92.87	79.21	66.11	69.62	77.61	57.91	62.78	**91.99**	78.37	82.98	**97.92**	34.67	45.62
	Open Source	Llama 3.3 70B	89.65	96.98	91.82	76.02	68.84	70.71	76.22	61.19	64.66	83.62	73.43	75.58	95.67	49.71	58.70
		Mistral Large	93.60	96.01	92.66	**88.07**	77.15	80.21	**86.60**	64.80	70.24	80.97	73.96	75.40	93.65	47.27	56.80
	Baselines	Şahinüç et al. (2024)	88.69	97.41	92.85	75.17	75.60	75.38	43.57	36.65	39.81	87.42	81.19	**84.19**	90.84	70.22	79.21
		Kardas et al. (2020)	78.06	79.69	76.37	-	-	-	56.62	44.89	46.65	50.91	66.10	53.20	74.40	83.24	77.68
Open Domain	**Proprietary**	GPT-4.1	57.75	53.47	54.23	70.60	61.91	64.26	65.11	49.79	53.86	83.60	76.94	78.74	93.95	71.19	77.13
		GPT-4o	90.70	88.14	87.67	76.51	65.64	69.44	78.29	56.23	62.04	**90.47**	72.95	78.29	97.24	24.06	34.99
		o4-mini	89.53	96.47	**91.19**	78.54	**72.89**	**73.94**	76.68	58.69	62.87	80.51	75.68	75.90	94.93	71.46	76.41
		Gemini-2.5-Pro	77.56	80.45	76.56	72.06	67.48	68.38	69.50	63.83	64.62	64.83	66.14	63.78	87.39	**98.30**	**91.87**
		Gemini-2.5-Flash	85.99	**98.50**	89.84	48.80	56.37	50.59	46.68	43.76	44.10	74.21	**78.72**	75.43	88.04	93.33	89.26
		Gemini-1.5-Pro	78.57	83.42	78.59	28.91	20.76	22.75	29.07	15.97	18.51	90.23	75.35	**80.01**	**98.19**	33.63	44.66
	Open Source	Llama 3.3 70B	88.93	95.91	90.64	75.77	67.03	69.43	**82.98**	**64.80**	**69.35**	87.27	75.22	78.64	92.62	47.22	56.32
		Mistral Large	**93.22**	89.22	88.72	**84.14**	68.20	72.87	78.96	58.99	63.73	83.10	74.01	76.83	95.11	50.65	60.50
	Baselines	Şahinüç et al. (2024)	13.82	17.51	15.49	21.11	24.31	22.84	25.27	25.04	25.12	61.79	59.68	60.73	90.84	70.22	79.32

Table 4: Model performance on individual entities in the METALEAD dataset for **open** and **closed** domains, with top closed-domain results in **blue** and open-domain results in **orange**.

	Model					
	Type	Name	LR	PC	RC	AO
Closed Domain	**Proprietary**	GPT-4.1	36.10	27.57	20.32	6.61
		GPT-4o	20.43	12.95	5.83	3.03
		o4-mini	30.11	22.80	14.97	5.95
		Gemini-2.5-Pro	**49.46**	**41.05**	**36.27**	**7.99**
		Gemini-2.5-Flash	23.96	13.98	11.79	**7.99**
		Gemini-1.5-Pro	16.44	10.82	6.07	2.76
	Open Weigths	Llama 3.3 70B	23.81	16.34	10.42	3.38
		Mistral Large	23.66	16.98	8.78	3.46
	Baselines	Şahinüç et al. (2024)	18.59	11.96	6.24	1.84
		Kardas et al. (2020)	14.50	6.99	5.95	0.00
Open Domain	**Proprietary**	GPT-4.1	7.07	5.47	3.42	0.92
		GPT-4o	13.67	8.74	4.43	2.00
		o4-mini	**25.19**	**17.95**	**12.80**	**3.53**
		Gemini-2.5-Pro	19.20	12.04	11.05	0.77
		Gemini-2.5-Flash	21.04	13.83	10.88	1.69
		Gemini-1.5-Pro	3.38	2.11	1.14	0.15
	Open Source	Llama 3.3 70B	21.66	15.07	9.20	2.61
		Mistral Large	3.38	2.11	1.14	3.00
	Baselines	Şahinüç et al. (2024)	3.07	1.59	0.84	0.15

Table 5: Leaderboard Construction Results with leaderboard recall (LR), paper coverage (PC), results coverage (RC), and, average overlap (AO), with top closed-domain results in **blue** and open-domain results in **orange**.

	Model		
	Type	Name	ET-Accuracy
Closed Domain	**Proprietary**	GPT-4.1	70.88
		GPT-4o	58.13
		o4-mini	68.31
		Gemini-2.5-Pro	72.78
		Gemini-2.5-Flash	47.05
		Gemini-1.5-Pro	59.93
	Open Source	Llama 3.3 70B	56.23
		Mistral Large	64.17
	Baselines	Şahinüç et al. (2024)	**81.62**
Open Domain	**Proprietary**	GPT4.1	31.46
		GPT-4o	62.53
		o4-mini	63.72
		Gemini-2.5-Pro	53.27
		Gemini-2.5-Flash	50.24
		Gemini-1.5-Pro	33.85
	Open Source	Llama 3.3 70B	59.87
		Mistral Large	62.96
	Baselines	Şahinüç et al. (2024)	**81.25**

Table 6: *Experiment type* classification accuracy (**ET-Accuracy**), with top closed-domain results in **blue** and open-domain results in **orange**.

score, and experiment type from the METALEAD dataset. In addition, future work could explore prompt engineering to better align the model with annotation guidelines (Sainz et al.; Huang et al., 2025). Vision-language models (VLMs), such as *Nanonets* (Mandal et al., 2025), may enhance extraction from tables and figures that are difficult for text-only models. Multi-agent systems also show promise (Song et al., 2025), by dividing extraction into specialised subtasks (e.g., tuple extraction, validation, experiment-type labelling) and cross-verifying outputs for greater consistency and completeness.

Flexible Leaderboards Traditional leaderboard schemas have been static, relying on fixed structures. Ideally, users should define their own schema and filter results based on individual needs, for ex-

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	89.67	0.66	9.68
Proposed Method	3.30	73.13	23.57
Method Variation	11.39	11.55	77.06

Table 7: Confusion matrix for *experiment type* classification by GPT-4.1 in the closed domain. Percentages are averaged over 39 papers with non-empty matrices.

ample, viewing only *ID* results, including *OOD* results, or selecting the best result and baseline per paper. This flexibility would make leaderboards more user-centric. Exploring how such dynamic, customisable leaderboards could work in practice presents an interesting direction for future research.

Benchmarking Negative Results and Model Failures METALEAD includes negative results and failed variants, traditionally excluded from traditional leaderboards. This enables systematic analysis of model weaknesses and failure patterns. Future work could develop meta-analysis tools to identify common conditions for poor performance, such as specific task types, dataset setups, or metrics. Highlighting and categorising failures promotes transparent reporting and supports the development of more robust models and evaluation frameworks (Karl et al., 2024).

Downstream Tasks METALEAD can be used for a wide range of downstream tasks. The simplest is for a researcher to quickly survey a field, identify gaps, and plan new experiments. More advanced applications include result-centric summarisation, citation recommendation, and trend analysis of datasets and metrics. METALEAD also supports predictive modelling, benchmark coverage auditing, and the construction of structured knowledge graphs (e.g. Kabongo et al. (2024b)). An important downstream task is model selection for large-scale industry deployment. Traditional leaderboards would only display the best result per paper (e.g. 91.62 and 90.32 F1 for NER in Figure 2), but an industry practitioner may also value robustness and cost. Using leaderboard that conforms to the METALEAD data schema, one may notice that a baseline method performs slightly below the top results(e.g., 90.20 F1 for NER). Such information is useful to the industry practitioner, as the baseline method may be cheaper, making it more cost-effective in a production setting. By including all experimental results with rich metadata, METALEAD can facilitate more transparent, scalable, and automated research workflows.

8 Conclusion

In this paper, we present METALEAD, a fully human-annotated comprehensive dataset designed to improve the transparency and utility of ML leaderboard reporting. METALEAD addresses key limitations of existing datasets by (1) including **all reported results**, such as baselines and ablations, rather than only the best ones; (2) categorising each result by **experiment type** to support structured comparisons; and (3) **distinguishing train and test datasets** to enable cross-domain assessment. These schema innovations promote more nuanced evaluations, reduce ambiguity, and facilitate downstream tasks such as benchmarking model robustness, auditing evaluation practices, and enabling user-defined leaderboard construction. As demonstrated by the benchmarking results, extracting complete and structured information remains challenging, but METALEAD provides a clear reference standard and supports the development of more rigorous and transparent ML research.

Limitations

Our work presents several important limitations. First, although METALEAD contains 3,568 annotated results, an order of magnitude more than SciLead, the number of source papers remains modest. Future extensions should expand beyond 43 ML papers to include other scientific domains and non-English publications, in order to enhance generalisability and coverage. Second, the papers are restricted to the machine learning domain, and the timeframe considered further narrows the applicability of the results. Third, we focus exclusively on papers written in English, which limits the applicability of our methods to non-English scientific literature. Fourth, while a leaderboard could theoretically be expanded to finer levels of granularity, such as more detailed score breakdowns or nuanced experimental conditions, we chose not to pursue this direction. This decision was informed by prior research that demonstrates low inter-annotator agreement when annotating machine learning scores and other experimental details within papers (Otto et al., 2023). Maintaining consistency and ensuring meaningful comparisons across such fine-grained annotations introduces significant challenges. As such, we recommend that researchers refer to the original

papers for complete experimental details. Finally, our extraction methods target experimental results presented in the main body of papers. We excluded results presented in figures (e.g. line graphs), as the corresponding values are often ambiguous or too difficult to extract reliably.

Ethical Considerations

For our annotations, we followed the instructions from the ethics board of the Commonwealth Scientific and Industrial Research Organisation (CSIRO). The annotations were conducted by the authors as part of their paid research duties. The authors, as domain experts familiar with the METALEAD papers, served as annotators, a common practice in leaderboard research. The dataset does not contain any sensitive information, and no personal information about the annotators was recorded. The METALEAD dataset will be made publicly available for research purposes. Given its focus on scientific literature and the nature of its content, we do not anticipate any risk of misuse for malicious purposes.

METALEAD is built upon the dataset introduced by Şahinüç et al. (2024), which was published under the *CC-BY-NC 4.0 (Attribution-NonCommercial 4.0 International)* licence. In accordance with the licence terms, we attribute the original authors and ensure that our use remains strictly non-commercial. The original dataset served as a foundation, which we substantially expanded through additional manual annotations and revised schema definitions. The combined dataset, including our extensions, will be released solely for academic research purposes. No modifications were made to the original data beyond reformatting for consistency, and the provenance of the original annotations is preserved.

References

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. 2018. Construction of the Literature Graph in Semantic Scholar. In *Proceedings of NAACL-HLT*, pages 84–91.

arXiv. 2025. arXiv Monthly Submissions Statistics. Accessed: 2025-01-20.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. 2022. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, 35:19274–19289.

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jennifer D'Souza, Salomon Kabongo, Hamed Babaei Giglou, and Yue Zhang. 2024. Overview of the CLEF 2024 simpletext task 4: SOTA? tracking the state-of-the-art in scholarly publications. *Working Notes of CLEF*.

Elicit. 2023. Elicit: The AI Research Assistant.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the Eye of the User: A Critique of NLP Leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. 2025. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*.

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.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2021. TDMSci: A Specialized Corpus for Scientific Literature Entity Tagging of Tasks Datasets and Metrics. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 707–714, Online. Association for Computational Linguistics.

Shizhou Huang, Bo Xu, Yang Yu, Changqun Li, and Xin Alex Lin. 2025. GuideNER: Annotation Guidelines Are Better than Examples for In-Context Named Entity Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24159–24166.

Hugging Face. 2024. Mistral Large Model Card. Accessed: 2025-07-24.

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.

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.

Salomon Kabongo, Jennifer D'Souza, and Sören Auer. 2024a. Effective context selection in llm-based leaderboard generation: An empirical study. In *International Conference on Applications of Natural Language to Information Systems*, pages 150–160. Springer.

Salomon Kabongo, Jennifer D'Souza, and Sören Auer. 2024b. ORKG-Leaderboards: a systematic workflow for mining leaderboards as a knowledge graph. *International Journal on Digital Libraries*, 25(1):41–54.

Salomon Kabongo, Jennifer D'Souza, and Sören Auer. 2021. Automated Mining of Leaderboards for Empirical AI Research. volume 13133 LNCS.

Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. 2020. AxCell: Automatic extraction of results from machine learning papers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8580–8594, Online. Association for Computational Linguistics.

Florian Karl, Lukas Malte Kemeter, Gabriel Dax, and Paulina Sierak. 2024. Position: embracing negative results in machine learning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 23256–23265.

Fech Scen Khoo, Megan Mark, Roelien C. Timmer, Marcella Scoczynski Ribeiro Martins, Emily Foshee, Kaylin Bugbee, Gregory Renard, and Anamaria Berea. 2023. Building knowledge graphs in heliophysics and astrophysics. In *Natural Language Processing and Information Systems*, pages 215–228, Cham. Springer Nature Switzerland.

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

Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. 2019. A Dual-Attention Hierarchical Recurrent Neural Network for Dialogue Act Classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 383–392.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021. ExplainaBoard: An explainable leaderboard for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 280–289, Online. 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. pages 3219–3232.

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.

Nick McGreivy and Ammar Hakim. 2024. Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations. *Nature machine intelligence*, 6(10):1256–1269.

Ishani Mondal, Yufang Hou, and Charles Jochim. 2021. End-to-end construction of NLP knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1885–1895.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

OpenAI. 2024a. GPT-4.1. `https://openai.com/index/gpt-4-1/`. Accessed: 2025-07-24.

OpenAI. 2024b. GPT-4o System Card. Technical Report arXiv:2410.21276, OpenAI, San Francisco, CA. Accessed: 2025-07-24.

OpenAI. 2025. Deep Research. https://openai.com/index/introducing-deep-research/. AI research assistant.

OpenAI. 2025. Introducing OpenAI o3 and o4-mini. Technical report, OpenAI. Accessed: 2025-07-24.

Wolfgang Otto, Matthäus Zloch, Lu Gan, Saurav Karmakar, and Stefan Dietze. 2023. GSAP-NER: A Novel Task, Corpus, and Baseline for Scholarly Entity Extraction Focused on Machine Learning Models and Datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8166–8176.

Evan F Risko and Sam J Gilbert. 2016. Cognitive offloading. *Trends in cognitive sciences*, 20(9):676–688.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction. In *The Twelfth International Conference on Learning Representations*.

Mayank Singh, Rajdeep Sarkar, Atharva Vyas, Pawan Goyal, Animesh Mukherjee, and Soumen Chakrabarti. 2019. Automated early leaderboard generation from comparative tables. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41*, pages 244–257. Springer.

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.

Zihan Song, Gyo-Yeob Hwang, Xin Zhang, Shan Huang, and Byung-Kwon Park. 2025. A scientific-article key-insight extraction system based on multi-actor of fine-tuned open-source large language models. *Scientific Reports*, 15(1):1608.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton Van Den Hengel. 2020. On the value of out-of-distribution testing: An example of Goodhart's Law. *Advances in neural information processing systems*, 33:407–417.

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 30761–30784, Suzhou, China. Association for Computational Linguistics.

Roelien C. Timmer, Megan Mark, Fech Scen Khoo, Marcella Scoczynski Ribeiro Martins, Anamaria Berea, Gregory Renard, and Kaylin Bugbee. 2023. Nasa science mission directorate knowledge graph discovery. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 795–799, New York, NY, USA. Association for Computing Machinery.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.

Sean Yang, Chris Tensmeyer, and Curtis Wigington. 2022. TELIN: Table Entity LINker for Extracting Leaderboards from Machine Learning Publications.

Fatih Şahinüç, Thi Thao Tran, Yuliya Grishina, Yufang Hou, Bowen Chen, and Iryna Gurevych. 2024. Efficient Performance Tracking: Leveraging Large Language Models for Automated Construction of Scientific Leaderboards. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Miami, Florida, USA. Association for Computational Linguistics.

A Annotation Guidelines

Objective: Identify and extract all experimental results from the paper, including baseline and ablation studies for each specified task. Extract results in the format *<Task, Train Dataset, Test Dataset, Metric, Result>*, e.g.: *<"Task": "Language Modeling", "Train Dataset": "WikiText-103", "Test Dataset": "WikiText-103", "Metric": "Perplexity", "Result": 28.0>* or *<"Task": "Language Modeling", "Train Dataset": "Penn Treebank (PTB)", "Test Dataset": "Penn Treebank (PTB)", "Metric": "Perplexity", "Result": 46.01>*

Guidelines:

- **Limit to the tasks provided:** Only report results related to the tasks in the original annotation (no other tasks). However, include results for all datasets and metrics associated with the task, even if the original annotations do not.
- **Avoid duplicates:** If the same result is reported multiple times (e.g., in multiple tables), ensure that you only report it once.
- **Limit to ML experiments:** Exclude experimental results from human annotators.
- **Differentiate between train and test datasets:** The training dataset and the test dataset are the same; however, in rare cases, they might differ.

Figure 4: Instructions Part 1 of the Annotation Process.

Objective: Label each experimental result based on its type of experiment.

Guidelines: Choose one of the following labels:

1. **Baseline comparison:** Results that are benchmarks or comparisons with prior work (e.g. previous state-of-the-art results).
2. **Proposed method:** Experiments with the main method proposed by the authors. The method proposed by the authors is generally regarded as their main contribution. Authors often refer to this method as "Our Model" or "Our Method". If several fine-tuning variants are tested (typically on a development set), the best-performing version that is used for test set evaluation should be labelled as the proposed method, while the others are considered ablations.
3. **Ablation study:** Experiments that test the contribution of specific features, modules, hyperparameter tuning, etc. If there are ablation studies for the baselines, label them as baseline comparisons, NOT ablation studies. Any hybrid/ensemble method that includes the proposed method should also be labelled as an ablation study. Experiments using the main proposed method, even when conducted on a subset of the data, should be labelled as *proposed method* and should not be classified as ablation studies, unless they do experiments on the subset of the data to determine best features, hyperparameter, etc. Changes to the pre-training setup (but not the train dataset) should also be treated as ablation studies.

If the authors clearly specify the type of experiment, follow their classification, otherwise, use your own judgement.

Figure 5: Instructions Part 2 of the Annotation Process.

B Prompts

Prompt for the first extraction step of the leaderboard tuples is shown in Figure 6, and, the normalisation prompt is shown in Figure 7.

```
You are given several parts of a research paper as
input.
Extract these tuples for all experimental results.
Each tuple must contain:

- a single task, single train dataset, single test
  dataset, single metric, single numeric result,
  experiment-type label.

Ensure tuples are not aggregated and contain a
single value for each field.
Use the following JSON format:
  {
  ''Task": ''Task name",
  ''Train-Dataset": ''Train dataset name",
  ''Test-Dataset": ''Test dataset name",
  ''Metric": ''Metric name",
  ''Result": ''Result score",
  ''Experiment-Type": ''Experiment type label
  (''proposed method'', ''baseline'', or
  ''variation of proposed method'')"
  }
```

Figure 6: Prompt for the first extraction step.

```
Match the input {input_item} with the best
candidate from {candidates_list}.
Return the exact matching item or the input if no
match is found.
Only return the exact matching item, nothing else.
```

Figure 7: Prompt for the normalisation step.

C Experiment Setup Details

For the proprietary and open source models, we have only selected models with a minimum of 5,000 output token. We use three OpenAI models with the Azure API: *GPT-4.1*, *GPT-4o*, and, *o4-mini*. The approximate cost for all the experiments with the Azure API is approximately USD 25 Of Google, we use three different versions of the Gemini family *Gemini-2.5-Pro*, *Gemini-2.5-Flash*, and, *Gemini-1.5-Pro*. The approximate cost for the Gemini API is USD 50. For Mistral, we use *Mistral-Large-2402*, for which the cost is approximately USD10. For Llama of Meta, we use the Huggingface version of *Llama 3.3 70B*. The experiments for *Llama 3.3 70B* were conducted locally on four Nvidia H100 GPUs. Before passing content to the large language models, all PDFs are parsed using PyMuPDF[13], which preserves the layout structure. We exclude the reference list and the appendix.

Responsible AI Content Filtering Some of the normalisation prompts triggered Azure's content filter, despite containing no harmful or inappropriate material. For instance, prompts that included the metric *V-M* consistently triggered Azure's content filter. As a result, we were required to slightly modify certain prompts to bypass these restrictions. When a prompt triggered the filter, we retained the originally extracted entity without applying further normalisation.

AxCell The AxCell baseline of Kardas et al. (2020), relies on input files for which source code is available. Consequently, AxCell can only be executed for 36 out of the 43 papers. Due to the way AxCell is originally designed, it can only extract tuples in the format ⟨*task, dataset, metric, score*⟩. For comparison with the METALEAD dataset, we align AxCell's extracted "dataset" field with the "test dataset" field in METALEAD. During the normalisation step, we employ the GPT-4.1 model. We have modified the original AxCell codebase to extract all available results, including baselines, ablations, and results with low scores, rather than filtering them out.

SciLead For the SciLead baseline of (Şahinüç et al., 2024), we use the GPT-4.1 model for two reasons. First, our own simple prompting experiments with GPT-4.1 yield together with Gemini-2.5-Pro the best performance. Second, in the study of Şahinüç et al. (2024), the implementation using a GPT-family model achieved the strongest performance.

D Experiment-Type Classification Confusion Matrices

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	81.62	10.00	8.38
Proposed Method	0.39	88.14	11.47
Method Variation	0.00	35.71	64.29

Table 8: Confusion matrix for *experiment type* classification by GPT-4o in the closed domain. Percentages are averaged over 34 papers with non-empty matrices.

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	97.47	0.70	1.83
Proposed Method	0.89	70.18	28.93
Method Variation	12.89	6.52	80.59

Table 9: Confusion matrix for *experiment type* classification by o4-mini in the closed domain. Percentages are averaged over 36 papers with non-empty matrices.

[13] https://github.com/pymupdf/PyMuPDF

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	94.70	0.47	4.83
Proposed Method	0.32	77.53	22.15
Method Variation	7.41	10.20	82.39

Table 10: Confusion matrix for *experiment type* classification by Gemini-2.5-Pro in the closed domain. Percentages are averaged over 37 papers with non-empty matrices.

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	95.83	0.76	3.41
Proposed Method	2.50	80.00	17.50
Method Variation	14.29	0.00	85.71

Table 11: Confusion matrix for *experiment type* classification by Gemini-2.5-Flash in the closed domain. Percentages are averaged over 23 papers with non-empty matrices.

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	87.33	11.00	1.67
Proposed Method	5.28	71.43	23.29
Method Variation	28.57	13.10	58.33

Table 12: Confusion matrix for *experiment type* classification by Gemini-1.5-Pro in the closed domain. Percentages are averaged over 35 papers with non-empty matrices.

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	81.04	4.72	14.24
Proposed Method	1.72	87.35	10.92
Method Variation	11.33	12.02	76.64

Table 13: Confusion matrix for *experiment type* classification by LLaMA 3.3 70B in the closed domain. Percentages are averaged over 32 papers with non-empty matrices.

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	85.89	5.59	8.52
Proposed Method	0.00	77.51	22.49
Method Variation	15.28	16.67	68.06

Table 14: Confusion matrix for *experiment type* classification by Mistral Large in the closed domain. Percentages are averaged over 37 papers with non-empty matrices.

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	67.79	2.30	11.73
Proposed Method	4.22	58.50	14.55
Method Variation	4.70	5.56	53.38

Table 15: Confusion matrix for *experiment type* classification by Şahinüç et al. (2024) in the closed domain. Percentages are averaged over 22 papers with non-empty matrices.

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	90.05	1.07	8.88
Proposed Method	1.33	71.07	27.59
Method Variation	7.29	6.85	85.86

Table 16: Confusion matrix for *experiment type* classification by GPT-4.1 in the open domain. Percentages are averaged over 18 papers with non-empty matrices.

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	90.00	4.17	5.83
Proposed Method	0.00	91.67	8.33
Method Variation	16.67	33.33	50.00

Table 17: Confusion matrix for *experiment type* classification by GPT-4o in the open domain. Percentages are averaged over 34 papers with non-empty matrices.

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	96.75	0.48	2.77
Proposed Method	0.59	66.00	33.41
Method Variation	12.50	9.38	78.13

Table 18: Confusion matrix for *experiment type* classification by o4-mini in the open domain. Percentages are averaged over 35 papers with non-empty matrices.

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	96.48	0.49	3.04
Proposed Method	1.14	90.44	8.42
Method Variation	7.14	2.34	90.52

Table 19: Confusion matrix for *experiment type* classification by Gemini-2.5-Pro in the open domain. Percentages are averaged over 25 papers with non-empty matrices.

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	97.94	0.00	2.06
Proposed Method	0.00	75.10	24.90
Method Variation	9.98	10.83	79.19

Table 20: Confusion matrix for *experiment type* classification by Gemini-2.5-Flash in the open domain. Percentages are averaged over 25 papers with non-empty matrices.

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	95.83	1.14	3.03
Proposed Method	6.25	83.33	10.42
Method Variation	33.33	5.56	61.11

Table 21: Confusion matrix for *experiment type* classification by Gemini-1.5-Pro in the open domain. Percentages are averaged over 17 papers with non-empty matrices.

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	85.61	5.17	9.22
Proposed Method	1.57	86.61	11.82
Method Variation	10.50	15.49	74.00

Table 22: Confusion matrix for *experiment type* classification by LLaMA-3.3-70B-Instruct in the open domain. Percentages are averaged over 35 papers with non-empty matrices.

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	79.82	4.94	15.24
Proposed Method	0.83	67.42	31.75
Method Variation	12.94	16.43	70.63

Table 23: Confusion matrix for *experiment type* classification by Mistral Large in the open domain. Percentages are averaged over 38 papers with non-empty matrices.

Gold / Predicted	Baseline	Proposed Method	Method Variation
Baseline	88.89	2.78	8.33
Proposed Method	12.50	75.00	12.50
Method Variation	6.25	0.00	18.75

Table 24: Confusion matrix for *experiment type* classification by Şahinüç et al. (2024) in the open domain. Percentages are averaged over 4 papers with non-empty matrices.

E Unique Entities in the METALEAD

'AVG', 'Accuracy', 'BERTScore', 'BLEU', 'BLEU-4', 'Exact Match (EM)', 'F-Score (F-S)', 'Fuzzy B-Cubed (FBC)', 'Fuzzy normalized mutual information (FNMI)', 'Labeled Attachment Score', 'METEOR', 'Matthew's Correlation Coefficient (MCC)', 'NIST-4', 'Overall-Accuracy', 'Perplexity', 'Precision', 'ROUGE-1', 'ROUGE-2', 'ROUGE-L', 'Recall', 'Sent-Accuracy', 'Spearman Correlation', 'TER', 'Unlabeled Attachment Score', 'V-Measure (V-M)'

Figure 8: The unique metrics in METALEAD

The 23 unique tasks in METALEAD are: *'Dialogue Generation', 'Text Chunking', 'Combinatory Categorial Grammar (CCG) Supertagging', 'Part-of-Speech (POS) Tagging', 'Question Answering', 'Relation Classification', 'Text Similarity', 'Dependency Parsing', 'Named Entity Recognition (NER)', 'Constituency Parsing', 'Intent Detection and Slot Filling', 'Question Generation', 'Paraphrase Detection', 'Entity Typing', 'Natural Language Inference (NLI)', 'Sentiment Analysis', 'Summarization', 'Response Generation', 'Machine Translation', 'Dialogue Act Classification', 'Word Sense Induction', 'Linguistic Acceptability', 'Language Modeling'.*

Figure 9: The unique tasks in METALEAD

'AESLC', 'ATIS', 'AX', 'BC5CDR', 'BIGPATENT', 'BillSum', 'CBS SciTech News', 'CCGBank', 'CCGBank Dev', 'CNN/DailyMail', 'CNN/DailyMail Dev', 'CoLA', 'CoLA Dev', 'CoNLL++', 'CoNLL-03 - English', 'CoNLL-2000', 'CoNLL-2000 Dev', 'CoNLL-2002 - Dutch', 'CoNLL-2002 - Dutch Dev IV', 'CoNLL-2002 - Dutch Dev OOBV', 'CoNLL-2002 - Dutch Dev OOEV', 'CoNLL-2002 - Dutch Dev OOTV', 'CoNLL-2002 - Dutch Test IV', 'CoNLL-2002 - Dutch Test OOBV', 'CoNLL-2002 - Dutch Test OOEV', 'CoNLL-2002 - Dutch Test OOTV', 'CoNLL-2002 - Spanish', 'CoNLL-2002 - Spanish Dev IV', 'CoNLL-2002 - Spanish Dev OOBV', 'CoNLL-2002 - Spanish Dev OOEV', 'CoNLL-2002 - Spanish Dev OOTV', 'CoNLL-2002 - Spanish Test IV', 'CoNLL-2002 - Spanish Test OOBV', 'CoNLL-2002 - Spanish Test OOEV', 'CoNLL-2002 - Spanish Test OOTV', 'CoNLL-2003 - English', 'CoNLL-2003 - English Dev', 'CoNLL-2003 - English Dev IV', 'CoNLL-2003 - English Dev OOBV', 'CoNLL-2003 - English Dev OOEV', 'CoNLL-2003 - English Dev OOTV', 'CoNLL-2003 - English IV', 'CoNLL-2003 - English OOBV', 'CoNLL-2003 - English OOEV', 'CoNLL-2003 - English OOTV', 'CoNLL-2003 - English Test IV', 'CoNLL-2003 - English Test OOBV', 'CoNLL-2003 - English Test OOEV', 'CoNLL-2003 - English Test OOTV', 'CoNLL-2003 - English Train', 'CoNLL-2003 - German', 'CoNLL-2003 - German Dev IV', 'CoNLL-2003 - German Dev OOBV', 'CoNLL-2003 - German Dev OOEV', 'CoNLL-2003 - German Dev OOTV', 'CoNLL-2003 - German Test IV', 'CoNLL-2003 - German Test OOBV', 'CoNLL-2003 - German Test OOEV', 'CoNLL-2003 - German Test OOTV', 'CoQA Dev', 'Connl-2003 - English Dev', 'ConvAI2', 'ConvAI2 Dev', 'DSTC7', 'Dailydialog (DyDA)', 'E-commerce', 'ELI5', 'ELI5 Dev', 'Genia', 'Gigaword', 'Gigaword 10K examples', 'Google Billion Word', 'Google Billion Word Valid', 'ICSI Meeting Recorder Dialog Act Corpus (MRDA)', 'IWSLT'14 EN-DE', 'IWSLT'15 DE-EN', 'IWSLT'15 DE-EN Dev', 'IWSLT'15 EN-VI', 'MNLI-m', 'MNLI-m Dev', 'MNLI-mm', 'MRPC', 'MRPC Dev', 'MapTask', 'Meeting Recorder Dialog Act Corpus (MRDA)', 'Multi-News', 'Multi-domain Food', 'Multi-domain Home', 'Multi-domain Movie', 'Multimodal', 'NCBI', 'NEWSROOM', 'New York Times (NYT)', 'Newsroom', 'OConnl-2003 - English', 'OntoNotes v5 - English', 'OntoNotes v5 - English Dev', 'Ontonotes v4 - Chinese', 'Ontonotes v5 - Englis Dev', 'Ontonotes v5 - English', 'Ontonotes v5 - English BC', 'Ontonotes v5 - English BN', 'Ontonotes v5 - English Dev', 'Ontonotes v5 - English MZ', 'Ontonotes v5 - English NW', 'Ontonotes v5 - English TC', 'Ontonotes v5 - English WB', 'Open Entity', 'Penn Treebank (PTB)', 'Penn Treebank (PTB) Dev', 'Penn Treebank (PTB) Dev IV', 'Penn Treebank (PTB) Dev OOBV', 'Penn Treebank (PTB) Dev OOEV', 'Penn Treebank (PTB) Dev OOTV', 'Penn Treebank (PTB) IV', 'Penn Treebank (PTB) OOBV', 'Penn Treebank (PTB) OOEV', 'Penn Treebank (PTB) OOTV', 'Penn Treebank (PTB) Train', 'PubMed', 'QNLI', 'QNLI Dev', 'QQP', 'QQP Dev', 'RTE', 'RTE Dev', 'ReCoRD', 'ReCoRD dev', 'Reddit TIFU', 'Resume - Chinese', 'Ritter Twitter', 'SNIPS', 'SQuAD 1.1', 'SQuAD 1.1 Dev', 'SQuAD 1.1 dev', 'SQuAD 2.0', 'SQuAD 2.0 Dev', 'SST-2', 'SST-2 Dev', 'STS-B', 'STS-B Dev', 'SemEval 2010 Task 14', 'SemEval 2013 Task 13', 'SemEval 2013 Task 13 with data enrichment', 'Switchboard Dialog Act Corpus (SWDA)', 'TA-CRED', 'WMT'13 EN-DE', 'WMT'14 EN-DE', 'WMT'14 EN-FR', 'WMT'16 RO-EN', 'WMT'17 EN-ZH', 'WNLI', 'WNUT-16 - English', 'WNUT-17 - English', 'Weibo - Chinese', 'WikiHow', 'WikiText-103', 'WikiText-103 Dev', 'Wikitext-2', 'Wikitext-2 Dev', 'XSum', 'XSum Dev', 'arXiv'

Figure 10: The unique train and test datasets in METALEAD