GSAP-ERE: Fine-Grained Scholarly Entity and Relation Extraction Focused on Machine Learning

Wolfgang Otto[1,2], Lu Gan[1,2], Sharmila Upadhyaya[1], Saurav Karmakar[1], Stefan Dietze[1,2]

[1]GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
[2] Heinrich-Heine-University Düsseldorf, Germany
{firstname.lastname}@gesis.org

Abstract

Research in Machine Learning (ML) and AI evolves rapidly. Information Extraction (IE) from scientific publications enables to identify information about research concepts and resources on a large scale and therefore is a pathway to improve understanding and reproducibility of ML-related research. To training and testing of IE models focused on fine-grained information in ML-related research, e.g. method training and data usage, we introduce GSAP-ERE. It is a manually curated fine-grained dataset of mentions of 63K ML-related entities and 35K relations distributed across 10 entity types and 18 semantically categorized relation types annoated in the full text of 100 ML publications. We show that our dataset enables fine-tuned models to automatically extract ML-related information that facilitate knowledge graph (KG) construction from scholarly papers or monitoring of computational reproducibility of AI research at scale. Additionally, we use our dataset as a test suite to explore prompting strategies for IE using Large Language Models (LLM). We observe that the performance of state-of-the-art LLM prompting methods is largely outperformed by our best fine-tuned baseline model (NER: 80.6%, RE: 54.0% for the fine-tuned model vs. NER: 44.4%, RE: 10.1% for the LLM). This disparity of performance between supervised models and unsupervised usage of LLMs suggests datasets like GSAP-ERE are needed to advance research in the domain of scholarly information extraction.

Code — https://data.gesis.org/gsap/gsap-ere#code
Dataset — https://doi.org/10.60914/c4c1d-s0587
Extended version — https://arxiv.org/abs/2511.0941

1 Introduction

Recent studies such as Raff (2019) have documented the declining reproducibility of ML-related research. In addition, determining what models and datasets are state-of-the-art for specific tasks is a very effort-intensive and increasingly challenging process. These problems are elevated not only by poor benchmarking practices (Ferrari Dacrema, Cremonesi, and Jannach 2019) but also by the surging amount of ML-related research, where understanding the authority, adoption and quality of models and datasets and their interdependencies becomes increasingly hard. Different to other disciplines, prior studies in computer science underline that reproducibility problems are to a large extent caused by poor reporting practices in scholarly publications, where only 4% of the assessed ML-related publications could be reproduced when the original authors did not respond to clarification requests (Pineau et al. 2021).

Scholarly information extraction facilitates mining and understanding of vast amounts of scholarly literature (Luan et al. 2018, 2019; Jain et al. 2020; Pan et al. 2023; Zhang et al. 2024) and usually involves Named Entity Recognition (NER) and Relation Extraction (RE), enabling the identification of entities and their relationships. The joined Entity and Relation Extraction (ERE) enables tasks, such as academic question answering (Dasigi et al. 2021) and scholarly knowledge graph construction (Viswanathan, Neubig, and Liu 2021), and thus facilitates reproducibility and reusability of research.

In particular among scholarly entities relevant to AI and ML-related research, e.g., dataset, ML model and task, various dependencies are of great interest to enhance research reproducibility and reusability (Arvan, Pina, and Parde 2022; Wonsil et al. 2023). *Data-model dependencies* describe the datasets used for training and evaluating models; *model-task dependencies* indicate the downstream tasks applied to the trained or fine-tuned models; *inner-entity-type dependencies* capture comparability between scholarly resources or research concepts; and *entity-tracing dependencies* reflect the provenance or source of research artifacts.

However, many existing IE methods build upon datasets with a coarse-grained entity label set (Luan et al. 2018; Zhang et al. 2024) that are not able to distinguish fine-grain artifact metadata. Second, despite the huge progress in recent years, vanilla LLMs (Yang et al. 2025; Touvron et al. 2023) still lag behind largely on fine-grained domain-specific IE tasks compared to fine-tuned models (Zhang et al. 2024). Hence, it is not suitable to directly apply vanilla LLMs for a high-quality fine-grained scholarly IE. Third, despite of initial efforts to create a fine-grained scholarly NER dataset like GSAP-NER (Otto et al. 2023) or an RE dataset with wide relation coverage like SciER (Zhang et al. 2024), a comprehensive fine-grained entity and relation extraction dataset is still lacking.

We therefore introduce GSAP-ERE, a comprehensive

	ScienceIE	SemEval18 Task 7	SciERC	SciER	GSAP-ERE
Annotation Unit	♣	♠	♠	♥	♥
# Publications	500	500	500	106	100
# Entity Types	3	-	6	3	10
# Relation Types	2	6	7	9	18
# Entities	9,946	7,505	8,094	24,518	62,619
# Relations	672	1,583	4,648	12,083	35,302
# Entities in Relations	1,824 (18%)	3,166 (42%)	6,272 (77%)	17,148 (70%)	46,680 (75%)
# entity Sentences	3,319	2,403	2,551	7,722	20,020
# null Sentences	654	153	136	2	6185
# Relations/Pub.	3.1	3.3	9.3	114.0	353.0
# Sentences/Pub.	8.1	5.1	5.4	72.9	262.0

Table 1: Comparison of GSAP-ERE and four datasets supporting NER and RE in scientific text. Note that our dataset is the biggest dataset in terms of annotated entities and relations so far and includes checked *null Sentences* without entity mentions as negative samples. Annotation units: ♣=Paragraph, ♠=Abstract, ♥=Full Text

dataset of fine-grained entity and relation types, based on high-quality expert annotations at span level focused on ten entity types important in ML research.
Our contributions include:

- A reusable data model that captures the complex relations among ML models, datasets and tasks. We introduce a set of 18 relation labels, which are systematically organized into seven distinct semantic relation groups. We provide the data model visualization in the extended version.
- We provide a dataset of more than 63K entity mentions and 35K fine-grain relation annotations in 100 scientific publications, encompassing connections among ML models, methods, datasets, and tasks. Our annotations further include additional links to references and URLs, and cover all sentences in the full text of the selected publications, including those without relations or entities defined by our data model (null cases).
- To demonstrate the utility of our dataset, we present fine-tuned baseline models that showcase the applicability of automatic entity and relation extraction using both pipeline and joint modeling approaches. We also report zero-shot and few-shot performance results for unsupervised prompting approaches with large language models (LLMs).

The dataset, code for replicating the baseline models, and the annotation guideline are provided on our project site https://data.gesis.org/gsap/gsap-ere.

2 Related Work

Scholarly IE Datasets Multiple datasets (Augenstein et al. 2017; Gábor et al. 2018; Luan et al. 2018; Schindler et al. 2021; Zhang et al. 2024), have been proposed as ground-truth datasets for scholarly information extraction tasks, including Named Entity Recognition (NER) and Relation Extraction (RE). We compare our GSAP-ERE with four well-established, manually annotated ground-truth datasets for scholarly IE in Table 1.

ScienceIE at SemEval 2017 (Augenstein et al. 2017) and SemEval 2018 Task 7 (Gábor et al. 2018) are two pioneering ground-truth datasets for scholarly IE that contain both entity and relation annotations. ScienceIE includes 500 paragraphs from open access journals and contains three entity types: *Task*, *Method* and *Material* and two relation types: *Hyponym-Of* and *Synonym-Of*. SemEval 2018 Task 7 provides six types of relations on annotation of entity span without typing based on abstracts from NLP publications. Another dataset SciERC (Luan et al. 2018) has been proposed, extending ScienceIE and SemEval 2018, where 500 publication abstracts from 12 AI conference or workshop proceedings are annotated. SciERC contains six entity types, including one type *Generic* supporting informal mentions of entities, and seven relation types. However, all these datasets are only based on abstracts (Gábor et al. 2018; Luan et al. 2018) or selected paragraphs (Augenstein et al. 2017) of publications, rather than full texts, which can cover more diverse linguistic styles (Zhang et al. 2024) and thus potentially render a wider range of entities and relations. Therefore, full-text-based ground-truth datasets are introduced. DMDD (Pan et al. 2023) is a dataset automatically annotated with distant supervision on the full text of 31 219 scientific articles. Yet DMDD does not contain relation annotations, hence is not suitable for the relation extraction task. SciREX (Jain et al. 2020) contains four entity types but its relation annotations are limited to clusters of mentions rather than individual entity mention pairs, overlooking the contextual information of each entity mention. SoMeSci (Schindler et al. 2021) contains the full text of 100 articles from PubMed Central Open Access subset; however it focuses on various software mentions rather than ML-related entity types. SciER (Zhang et al. 2024), deriving from the NER dataset SciDMT (Pan et al. 2024), provides a corpus of 106 full-text annotated publications with three entity types (*Dataset*, *Task*, *Method*) and ten relation types. To our best knowledge, SciER has the most comprehensive relation annotation scheme among the existing comparable datasets. However, it only includes explicitly named mentions, excluding the unnamed or informal mentions of en-

Semantic Group	Purpose	Example
Model Design	Captures internal structure and dependencies of methods and ML models.	*architecture*: **BART** follows a standard **sequence-to-sequence Transformer** architecture.
Task Binding	Links models and datasets to the associated tasks.	*appliedTo*: For **NER**, we consider **CoNLL-2002** (Sang, 2002) and CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) datasets . . .
Data Usage	Describes how datasets are used in methods and models.	*evaluatedOn*: To that end, we evaluate **XLM-R** on the **GLUE** benchmark.
Data Provenance	Tracks the origin and transformation of datasets.	*sourcedFrom*: **PubMed** alone has a total of **29M articles** as of January 2019.
Data Properties	Encodes metadata of datasets (size and instance type).	*size*: **GOOAQ** contains **3 million questions** with short, snippet, or collection answers, such as the ones shown in Fig. 1.
Peer Relations	Covers semantic, logical and structural relation among same entity type.	*isComparedTo*: Comparing to **WIKISQL-WEAK** , **WIKITABLE-QUESTIONS** requires more complicated reasoning capabilities.
Referencing	Refers to external sources or documentation (citation and URL).	*citation*: The latter is inspired by **TAPAS** (**Herzig et al., 2020**) . . .

Table 2: Brief definition of each semantic relation group and a corresponding example from our GSAP-ERE dataset. Note that each example shows only one relation annotation that represents the corresponding semantic group. Other entity mentions and relations in the gold data are not shown. The highlighted entity types (**MLModel** , **ModelArchitecture** , **Task** , **Dataset** , **DatasetGeneric** , **DataSource** , **ReferenceLink**) match the colors used in the examples.

tities, which accounts for a substantial amount of mentions in research publications, and results in an incomplete set of relation annotations. Additionally, the authors do only provide annotated sentences, resulting in only 73 sentences per document in the training and test data.

Compared to these datasets, our dataset, derived from GSAP-NER (Otto et al. 2023), is based on a manual annotation of 100 full-text research publications in the fields of ML and applied ML. It contains 10 fine-grained entity types and 18 relation types. The entity label set includes both named and informal entity types, in contrast to SciER. To enable a wide range of downstream use cases, the 18 relation types cover 7 different semantic groups relating to the creation, usage, property, application and referencing of scholarly entities. Furthermore, the co-existence of flat, nested and overlapping entity mentions renders it an even more challenging relation extraction task.

Entity and Relation Extraction Entity and relation extraction, aiming at extracting relations from text, is composed of two sub-tasks: NER and RE tasks. To achieve the end-to-end relation extraction, there exists two major paradigms: the pipeline-based approaches (Zhong and Chen 2021; Ye et al. 2022) and the joined extraction approaches (Luan et al. 2019; Yan et al. 2023). PURE (Zhong and Chen 2021) is an pipeline-based approach that uses two pre-trained encoders for the NER and RE models to capture the contextual information for each span or span pair. It also proposes an approximate batched relation model that shares the position embedding between the entity marker and the entity start and end tokens to accelerate the inference. `PL-Marker` (Ye et al. 2022) is another pipeline-based approach, extending PURE batched relation model with packed levitated marker to model entity pair interaction. `HGERE` (Yan et al. 2023) is a state-of-the-art joint extraction approach back-boned by the PL-Marker framework (Ye et al. 2022) and incorporates a hypergraph neural network to facilitate information flow among entities and relations. We fine-tune and evaluate both state-of-the-art pipeline and joint ERE models to compare their performance on GSAP-ERE.

3 Preliminaries

We formalize the entity and relation extraction (ERE) task at the text span level, enabling the tracing of extracted entities and relations back to the specific sentences from which they originate. Formally, we define two sub-tasks, NER and RE, as follows. We consider a collection of publications D as input. Each publication is divided into paragraphs $P = \{p_1, p_2, \dots, p_n\}$, with each paragraph further segmented into consecutive sentences $S = \{s_1, s_2, \dots, s_m\}$. For each tokenized sentence $s_i = \{w_1, w_2, \dots, w_n\}$, we define two extraction steps to identify entities and relations as follows.

NER We define the task of NER with a closed set of entity labels L^{NER} as the identification of contiguous word sequences $e_i = \{w_l, \ldots, w_r\}$ (i.e., spans) with a corresponding entity label from L^{NER}.

RE In the second step, we define a closed RE task using a label set L^{RE}, which includes a NIL label to denote the absence of a relation. For every possible pair of entities $E^p = \{(e_i, e_j)|i \neq j\}$ we define the task of RE as the assignment of all pairs to one of the labels L^{RE}.

ERE is the joined task of NER and RE from end-to-end independent of the design of the used approaches.

Metrics A joint evaluation, consistent with the task definition of ERE, encompasses both entity and relation checking. For NER task, exact match (***NER***) and partial match (***NER***≈) are two common evaluation settings. For RE, we note that its difficulty is substantially higher than that of NER. This is due to the fact that any relation pair includes two entity mentions with a relation label, where both entity mentions and relation label can lead to misalignment. Therefore, to support the assessment of relation annotations, we introduce four evaluation settings as below.

RE+ refers to both the correct labels for the entities and the relation, and exact span matching for the entities;

RE asks only correct relation labels and exact entity spans, with no restriction for the entity types;

RE+≈ requires exact match of all labels but allows partially overlap of entity spans;

RE≈ only asks the exact relation label and overlapped entity spans, with no further requirement.

We report micro F1 metrics, unless otherwise specified.

4 GSAP-ERE Dataset

In this section, we present the detailed curation process of our GSAP-ERE dataset. We reuse the comprehensive entity label sets of GSAP-NER (Otto et al. 2023) and their entity annotation as a starting point to create relation annotations.

Data Collection and Processing

Since we extend and build upon the GSAP-NER dataset (Otto et al. 2023) with relation annotation, we first give a brief overview of the GSAP-NER dataset. GSAP-NER provides a corpus of 100 manually annotated full-text research publications in ML or applied ML fields. The publications are first sampled using a popularity-diversity combined strategy described in Otto et al. (2023). The publications are selected from a Hugging Face model search (Hugging Face 2016) and arXiv metadata (Cornell University 2022), and then converted from PDF to plain text format via GROBID (Lopez 2008). The ten entity types in this corpus are categorized into three categories: (1) ML model related, i.e. *MLModel*, *ModelArchitecture*, *MLModelGeneric*, *Method* and *Task*; (2) dataset related, i.e. *Dataset*, *DatasetGeneric* and *DataSource*; (3) miscellaneous, including *ReferenceLink* and *URL*.

Data Annotation

Annotation Scheme We aim at a fine-grained relation label set that can reflect the comprehensive interactions among ML models, datasets, tasks, methods, model architectures and the miscellaneous entities. The interaction information should serve to improve research reproducibility and reusability, covering the provenance, usage, transformation, application and comparison of the scholarly entities. Unlike SciER (Zhang et al. 2024), we keep the informal mentions of named entities, corresponding to entity types *MLModelGeneric* and *DatasetGeneric*, in the annotation scheme following SciERC (Luan et al. 2018). We argue that the informal mentions not only capture valuable information for coreferential and compositional relations, but also increase the entity-per-sentence density to cover a larger amount of relation interactions and alleviate data sparsity for certain relation types. Therefore, we define a label set of 18 relation types for seven semantic groups.

We provide examples for each semantic group in Table 2 and list the domain and range information for each relation in our seven semantic groups in Table 3. Find examples for

Domain	Relation	Range
Model Design		
Method	usedFor	Method MLModel(Generic)
MLModel(Generic) Method	architecture	ModelArchitecture
MLModel(Generic)	isBasedOn	MLModel(Generic)
Task Binding		
MLModel(Generic) Method	appliedTo	Task
Dataset(Generic)	benchmarkFor	Task
Data Usage		
MLModel(Generic) Method	trainedOn evaluatedOn	Dataset(Generic)
Data Provenance		
Dataset(Generic)	transformedFrom	Dataset(Generic)
	generatedBy	Method
	sourcedFrom	DataSource
Data Properties		
Dataset(Generic)	size hasInstanceType	DatasetGeneric
Peer Relations		
Any	coreference isPartOf isHyponymOf isComparedTo	Same as Subject
Referencing		
Any	citation	ReferenceLink
	url	URL

Table 3: Domain and range information for all 18 relation types in our 7 semantic groups.

each relation in the extended version (Otto et al. 2025). More detailed definitions of relations are provided in the guideline on our project site.

Annotation Strategy To ensure high annotation quality, we employ a two-phase annotation-refinement process. Similar to SciER (Zhang et al. 2024) and GSAP-NER (Otto et al. 2023), we use the INCEpTION (Klie et al. 2018) platform.

During the initial annotation phase, annotations are conducted by two student annotators with computer science background, who have finished an annotation training. Among the 100 publications, we randomly selected 10 publications for double annotation by the two annotators separately; and the remaining 90 publications are split and assigned to a single annotator each. In this phase, the annotators annotate the relations on the assigned publications with existing entity annotations introduced by GSAP-NER (Otto et al. 2023). In addition, the annotators are asked to check the existing entity annotations first to ensure the relation annotations are based on correct entity annotations.

In the refinement phase, we have two PhD students and two postdoc researchers in computer science to inspect the annotation alignment of the two annotators. Misalignment cases are filtered out in the 10 commonly annotated documents by both annotators, and patterns of misalignment on each relation are extracted. Relations in jointly annotated documents that show such misalignment patterns are re-checked and corrected if necessary on all 100 publications.

Human Agreement and Quality Enhancement We calculate interrater agreement on the 10 jointly annotated publications. We note that relation annotation is highly influenced by the underlying entity annotation. A revision of entity annotations preceding the relation annotation improves the NER interrater agreement from reported 0.61 (Otto et al. 2023) to 0.82 macro-F1 score under *NER* setting and 0.69 to 0.86 under *NER*≈ setting. We report NER interrater agreement in the extended version. Furthermore, due to the increasing difficulty of relation annotation, we evaluate the relation annotation in four settings as described in Section 3. We show the interrater agreement per semantic group in Table 4. Despite achieving a good overall alignment, the various agreement levels in different semantic groups confirm the challenge of our relation annotation, which may be partially due to the poor reporting practices mentioned in Section 1.

Semantic Group	*RE+*	*RE*	*RE+*≈	*RE*≈
Model Design	38.4	40.0	45.9	49.3
Task Binding	40.6	45.6	44.1	50.0
Data Usage	61.3	62.8	64.8	67.5
Data Provenance	51.1	51.5	58.6	60.0
Data Properties	80.1	80.1	84.6	84.6
Peer Relation	57.1	61.4	59.3	65.2
Referencing	61.6	68.1	65.2	73.7
weighted all	53.7	58.1	56.9	62.8

Table 4: Interrater agreement across different relation groups measured by F1 scores (%, micro average) under different evaluation settings explained in Section 3.

5 Experiments

We use our dataset to evaluate model abilities to extract relational knowledge from scholarly publications according to our problem definition in Section 3.

Baselines

Supervised PLM-based Baselines Supervised entity and relation extraction (ERE) using pretrained language models (PLMs) is demonstrating state-of-the-art performance in the domain of scholarly documents (Zhang et al. 2024). In our experiments we compare a pipeline-based approach, i.e. a separate NER step followed by a RE classification of the extracted entities, to a joint approach which learns to predict and classify entity mention spans and relations between pairs of detected entity mentions simultaneously using a joint loss.

PL-Marker (Ye et al. 2022). This pipeline-based approach first performs a span classification (NER) task, and then uses the proposed Packed Levitated Marker to model and evaluates the interaction between entity mention pairs for RE.

HGERE (Yan et al. 2023). This state-of-the-art joint approach extends the PLMarker framework and incorporates a hypergraph neural network to facilitate a high-order span classification for NER and entity pair classification for RE in one step.

We fine-tune `PL-Marker` and `HGERE` on our training and validation sets and then report the performance of each fine-tuned models, denoted as $\texttt{PL-Marker}_{\text{GSAP-ERE}}$ and $\texttt{HGERE}_{\text{GSAP-ERE}}$, on our test set. More specifically, we introduce and optimize a loss weight scheme in the `HGERE` approach in combination with *ternary* information flow configuration for the Hypergraph Neural Network (HGNN). For $\texttt{HGERE}_{\text{GSAP-ERE}}$ we tuned the hyperparameters in a stepwise process. We first adjusted the learning rate and found that `2e-5` gave the best results (tested: `1e-4`, `5e-5`, `2e-5`, `1e-5`). After that, we varied the batch size and obtained the best value with `18` (tested: `10`, `14`, `18`, `22`). We then tested different loss weighting factors to balance the NER and relation losses. Since the NER part learned faster, we increased the relative weight of the relation loss using the scheme $(1 - L) \cdot \text{loss}_{\text{ner}} + L \cdot \text{loss}_{\text{rel}}$. The best setting was $L = 0.9$ (tested: 0.1, 0.25, 0.5, 0.75, 0.9, 0.95). Finally, we varied the number of epochs and chose `8` (tested: 6, 8, 10). More experimental settings can be found in the extended version (Otto et al. 2025).

LLM Prompting We run zero-shot and few-shot prompts on two state-of-the-art open-source generative LLMs, i.e. `LLaMA 3.1` (Touvron et al. 2023) and `Qwen 2.5` (Yang et al. 2025). Our prompting pipeline for NER and RE follows a two-stage process comparable to the supervised pipeline approach. For each input sentence, we first extract entities. These entities are then used to form subject-object candidate pairs, which are passed to a second prompt to clas-

ERE-method	Model	*NER*	*NER≈*	*RE*	*RE≈*	*RE+*	*RE+≈*
Supervised pipeline	PL-Marker$_{\mathrm{GSAP-ERE}}$	72.6	77.7	41.4	46.2	36.3	39.9
Supervised joined loss	HGERE$_{\mathrm{GSAP-ERE}}$	**80.6**	**85.8**	**54.0**	**59.8**	**46.9**	**51.3**
	Qwen 2.5 32b	42.0	56.9	7.2	14.6	7.2	10.9
Unsupervised (LLM) pipeline	Qwen 2.5 72b	**44.4**	**59.1**	**10.1**	**15.7**	**8.2**	**11.9**
	LLaMA 3.1 72b	40.5	55.0	6.4	9.6	5.7	7.8

Table 5: Comparison of test set F1 score (%, micro average) for our baseline approaches. Label-strict metrics are marked with "+"; span-partial metrics with "≈". LLM-based results for NER are obtained with $k = 10$ examples and RE with $k = 1$ example. For both tasks we used the best performing configurations using the *similar+diverse* example selection strategy.

sify their relation label. For n predicted entities, we generate $n(n-1)$ entity pairs and classify each using an individual prompt. Example prompts and templates used in the experiment are provided in the Appendix of the extended version (Otto et al. 2025). Each prompt is composed of five structured sections: *Task Introduction*, *Label Definitions*, *Step-by-Step Instructions*, *Few-Shot Examples* and *Main Input*, i.e. the actual input sentence to annotate. This structured prompt pattern is found to largely improve few-shot learning performance by helping align output generation with task expectations (Madaan and Yazdanbakhsh 2022).

To select the best performing hyperparameters, we explore two key prompt design parameters: the number of few-shot examples, and the example selection strategy. We apply various hyperparameter settings to the validation set to pick the best performing configuration before applying it to the test set. For the example selection, we compare random selection (a fixed, randomly chosen set) to a dynamic retrieval-based strategy. The dynamic strategy can be one of:

similar: retrieving top-k similar sentences via cosine similarity

similar+diverse: combining similar with maximizing unique label re-ranking

For our experiments, all examples were chosen from the training set. We evaluate the *random* and *similar+diverse* strategies on the ERE task. Each result was obtained from a single run per prompt.

Experimental Setup

For the supervised approaches, we fork the implementations of Ye et al. (2022) (code: Natural Language Processing Lab at Tsinghua University (2022)) and Yan et al. (2023) (code: Zhaohui Yan (2023)) and document the steps required to setup and train the models. For the supervised methods, we use the scibert-scivocab-uncased as encoder (Beltagy, Lo, and Cohan 2019) (model: AllenAI (2019)). All supervised models are tested and evaluated using 10% of the annotated publications as validation and 10% as test set leaving 80 annotated publications for training. We selected for the *ternary* information flow configuration for the Hypergraph Neural Network in `HGERE` through hyperparameter optimization. We also tuned learning rate, batch size, and loss-weighting. Details of the optimization processes are provided in our Appendix in Otto et al. (2025). We run our best performing model with 5 random seeds and report the mean performance including the standard deviation for this model.

To guarantee reproducibility for the unsupervised prompting approach, we set the temperature parameter for all LLMs to zero. To select the best performing hyperparameter for the unsupervised prompting approach, we evaluated hyperparameters using `Qwen2.5 32B` on the validation set. For the final model comparison, we used three open-source LLMs: `Qwen2.5 (32B and 72B)` (Yang et al. 2025), and `LLaMA 3.1 70B` (Touvron et al. 2023) on the test set. For efficient local inference, we accessed the quantized versions of `Qwen 2.5` (model : Ollama (2025b; 2025c)) and `LLaMA 3.1 70B` (model: Ollama (2025a)) using the Ollama (2024) framework. For few-shot sentence selection, we use `multi-qa-mpnet-base-cos-v1` (model: sentence-transformers (2029)) from the Sentence-Transformers library (Reimers and Gurevych 2019) for sentence cosine similarity.
Experiments were conducted on a server running Ubuntu 22.04.4 LTS with 2× Intel Xeon 2.1 GHz CPUs (48 cores, 96 threads), 1.4 TB RAM, and 8 GPUs (4× RTX 2080 Ti with 11 GB and 4× A40 with 48 GB). The system also includes a multi-tier storage setup with both SSD and HDD arrays, totaling over 40 TB.

Evaluation Metrics

To comprehensively evaluate the performance of our baselines, we employ a set of metrics that capture different aspects and strictness levels for NER and RE introduced in the Preliminaries section. In general, we report micro F1 performance to reflect overall model performance across labels (Harbecke et al. 2022). In specific evaluation settings, we report micro, macro, and weighted-averaged metrics to summarize performance for detailed comparability.

6 Experimental Results

Baseline Comparison

Table 5 presents a comparative overview of the supervised pipeline, joined, and the unsupervised prompting approaches using LLMs. Our best supervised approach outperforms the LLM-based approaches by a large margin with an *NER* F1 performance of $80.6 \pm 0.3\%$ and *RE+* F1 of $46.9 \pm 0.5\%$. We report the best results for each setting with

Example Selection		Metric	0-shot	1-shot	2-shot	5-shot	10-shot	20-shot
random	*NER*	micro-F1	19.1	24.7	23.1	29.7	34.1	**34.4**
		macro-F1	23.3	27.0	26.5	28.6	**31.5**	29.7
		weighted-F1	16.1	22.0	20.3	28.4	33.5	**33.6**
	NER≈	micro-F1	33.0	41.8	37.1	50.1	**53.3**	50.3
		macro-F1	34.5	40.3	37.7	43.7	**46.0**	40.9
		weighted-F1	29.8	39.6	34.5	49.4	**53.1**	49.4
similar+diverse	*NER*	micro-F1	19.1	34.7	38.2	40.4	**40.9**	27.8
		macro-F1	23.3	34.3	37.6	39.6	**39.7**	21.2
		weighted-F1	16.1	34.0	37.9	40.6	**41.4**	28.0
	NER≈	micro-F1	33.0	53.8	56.7	58.1	**58.4**	39.4
		macro-F1	34.5	48.1	50.9	**52.1**	52.0	28.4
		weighted-F1	29.8	53.4	56.5	58.2	**58.6**	38.8

Table 6: NER performance for different number of k-shot and example selection strategies on Qwen2.5-32B. We report micro, macro, and weighted F1 (%) for exact (NER) and partial match (NER≈) on the validation set.

optimized hyperparameters. Specifically, the LLM-based results are all based on *similar+diverse* strategy, with 10 examples for NER evaluations and 1 example for RE evaluations. We detail the hyperparameter selection experiments in Section 6.

We first observe a substantial performance advantage of supervised models over unsupervised LLM prompting approaches, ranging from 18.6–39.8% for ***NER*/*NER*≈** to 28.1%–50% for all four RE settings. Comparing the two supervised PLM approaches, the joint approach `HGERE` outperforms the pipeline approach `PL-Marker` under every setting for both NER and RE, demonstrating the robustness of this start-of-the-art method. Finally, among LLM prompting results, `Qwen 72b` achieves the best performance for all settings. However, compared to the two supervised PLM approaches, the advantage of `Qwen 72B` over other LLMs is marginal, in particular for the RE task, marking more research effort is needed for LLMs on such a comprehensive fine-grained IE task. Our findings are consistent with similar observations reported in related work (Zhang et al. 2024).
When comparing runtime between the supervised and unsupervised approaches, the PLM-based methods outperforms the LLM-based methods by a substantial margin. On the test corpus of 10 documents, inference with the LLM approach is 182 times slower than with the fine-tuned PLM on the same hardware described in Section 5 (4 minutes vs. 12 hours and 29 minutes), while training the PLM required 2 hours and 30 minutes for a single training run.

k-Shot	RE≈	RE+≈	RE	RE+
0	16.8	11.7	8.9	6.5
1	20.4	**14.4**	10.7	8.0
2	**20.5**	**14.4**	**10.8**	**8.1**
5	19.9	14.0	10.3	7.6

Table 7: F1 scores (%, micro average) for the RE task on the validation set using Qwen2.5-32B.

Impact of LLM Few-Shot Hyperparameters

For our unsupervised prompting approach we run our pipeline on the validation set under different configurations to pick the optimized number of few-shot examples (k) and example selection strategy. Table 6 shows the results on NER in these configurations.[1] We observe that, consistently, the best performing configuration for NER is $k = 10$ examples with *similar+diverse* example selection, achieving 58% F1 score, +5% compared to random example selection. The performance sharply declines after 20 examples. For RE, we run $k \in \{0, 1, 2, 5\}$ with *similar+diverse* strategy. As shown in Table 7, the overall RE performance is unsatisfying, but 1-shot and 2-shot achieve the best performance with minimal performance differences.Taking the trade-off between extraction performance and computational costs into consideration, we select 10-shot *similar+diverse* for NER and 1-shot *similar+diverse* for RE as the optimized hyperparameters. The detailed results on the test set for pipeline-based NER and RE are provided in the extended version.

7 Conclusion

We introduced GSAP-ERE, a manually annotated dataset designed for fine-grained scholarly information extraction in ML research. It features 10 detailed entity types and 18 relation types annotated across the full-text of scientific articles, comprising more than 62K annotated entities and 35K relations. Both data model and dataset shed light on the interdependencies between key concepts involved in machine learning-related research and facilitate research into information extraction methods that can provide a more structured view on machine learning-related artefacts and their relations at scale.

Our experiments show that fine-tuned models trained on this dataset significantly outperform prompting methods using LLMs, highlighting the current limitations of unsupervised approaches in domain-specific IE tasks. The sig-

[1]We include the detailed performance in our extended version for further fine grained analysis.

nificant performance gap supports our claim that curated datasets like GSAP-ERE are essential to advance reliable extraction of research metadata, enabling downstream applications such as knowledge graph construction or the understanding of reproducibility of ML-based research at scale.

8 Limitations

With respect to the coverage of the dataset, we include only publications from (applied) machine learning to simplify the annotation task and focus the required expertise for annotation.

We emphasize the complexity of RE annotation, and will transfer the data curation lessons when we extend the corpus with more diverse fields. In addition, our corpus covers only sentence-level annotations. Although our sentence-level ERE dataset is already challenging due to the multiplicity of entity mentions and relations, future work on a document-level corpus would help advancing the task further. Note that document-level relation extraction also needs document-level coreference resolution, and the coreference clusters may include unnamed generic mentions, which requires an additional annotation step.

Ethical Statement

All documents in this dataset come from public arXiv submissions. These documents are already publicly available under the terms set by arXiv. We removed personal data such as author email addresses and kept only the content needed for research on scientific text processing.

Human annotators were informed about the goals of the project and agreed to take part. No sensitive personal data was collected from annotators. The annotation process followed clear guidelines, and we monitored annotation quality throughout the project.

The model we present can extract entities and relations from scientific text. It is not meant for use on private or sensitive documents. Users should apply it only to text they are allowed to process.

We release the dataset and model under licenses that support research use. We provide documentation to help others understand the limits of the data, possible errors, and the expected behavior of the model.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. This work has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the Projects BERD@NFDI (grant number 460037581), NFDI4DS (grant number 460234259), as well as Unknown Data (grant number 460676019) and ConnOSS (grant number 561044496).

References

AllenAI. 2019. scibert_scivocab_uncased (Hugging Face Model Card). https://huggingface.co/allenai/scibert_scivocab_uncased. Accessed: 2025-12-16.

Arvan, M.; Pina, L.; and Parde, N. 2022. Reproducibility in Computational Linguistics: Is Source Code Enough? In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2350–2361. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Augenstein, I.; Das, M.; Riedel, S.; Vikraman, L.; and McCallum, A. 2017. SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In Bethard, S.; Carpuat, M.; Apidianaki, M.; Mohammad, S. M.; Cer, D.; and Jurgens, D., eds., *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 546–555. Vancouver, Canada: Association for Computational Linguistics.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620. Hong Kong, China: Association for Computational Linguistics.

Cornell University. 2022. arXiv Dataset (Kaggle), Version 123. https://kaggle.com/datasets/Cornell-University/arxiv/versions/123. Accessed: 2025-12-16.

Dasigi, P.; Lo, K.; Beltagy, I.; Cohan, A.; Smith, N. A.; and Gardner, M. 2021. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4599–4610. Online: Association for Computational Linguistics.

Ferrari Dacrema, M.; Cremonesi, P.; and Jannach, D. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys ’19, 101–109. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362436.

Gábor, K.; Buscaldi, D.; Schumann, A.-K.; QasemiZadeh, B.; Zargayouna, H.; and Charnois, T. 2018. SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers. In Apidianaki, M.; Mohammad, S. M.; May, J.; Shutova, E.; Bethard, S.; and Carpuat, M., eds., *Proceedings of the 12th International Workshop on Semantic Evaluation*, 679–688. New Orleans, Louisiana: Association for Computational Linguistics.

Harbecke, D.; Chen, Y.; Hennig, L.; and Alt, C. 2022. Why only Micro-F1? Class Weighting of Measures for Relation Classification. arXiv:2205.09460.

Hugging Face. 2016. Hugging Face Models (sorted by downloads). https://huggingface.co/models?sort=downloads. Accessed: 2022-11-13.

Jain, S.; van Zuylen, M.; Hajishirzi, H.; and Beltagy, I. 2020. SciREX: A Challenge Dataset for Document-Level Information Extraction. In *Proceedings of the 58th Annual Meet-*

ing of the Association for Computational Linguistics, 7506–7516. Online: Association for Computational Linguistics.

Klie, J.-C.; Bugert, M.; Boullosa, B.; de Castilho, R. E.; and Gurevych, I. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Lopez, P. 2008. GROBID (GitHub Repository). https://github.com/kermitt2/grobid. Accessed: 2025-12-16.

Luan, Y.; He, L.; Ostendorf, M.; and Hajishirzi, H. 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*, 3219–3232. Brussels, Belgium: Association for Computational Linguistics.

Luan, Y.; Wadden, D.; He, L.; Shah, A.; Ostendorf, M.; and Hajishirzi, H. 2019. A general framework for information extraction using dynamic span graphs. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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)*, 3036–3046. Minneapolis, Minnesota: Association for Computational Linguistics.

Madaan, A.; and Yazdanbakhsh, A. 2022. Text and Patterns: For Effective Chain of Thought, It Takes Two to Tango. arXiv:2209.07686.

Natural Language Processing Lab at Tsinghua University. 2022. PL-Marker (GitHub Repository). https://github.com/thunlp/PL-Marker. Accessed: 2025-12-16.

Ollama. 2024. Ollama: Run Large Language Models Locally. https://ollama.com/. Accessed: 2025-12-16.

Ollama. 2025a. llama3.1:70b-instruct-q4_K_M (Ollama Library Entry). https://ollama.com/library/llama3.1:70b-instruct-q4_K_M. Accessed: 2025-12-16.

Ollama. 2025b. qwen2.5:32b (Ollama Library Entry). https://ollama.com/library/qwen2.5:32b. Accessed: 2025-12-16.

Ollama. 2025c. qwen2.5:72b (Ollama Library Entry). https://ollama.com/library/qwen2.5:72b. Accessed: 2025-12-16.

Otto, W.; Gan, L.; Upadhyaya, S.; Karmakar, S.; and Dietze, S. 2025. GSAP-ERE: Fine-Grained Scholarly Entity and Relation Extraction Focused on Machine Learning. arXiv:2511.09411.

Otto, W.; Zloch, M.; Gan, L.; Karmakar, S.; and Dietze, S. 2023. GSAP-NER: A Novel Task, Corpus, and Baseline for Scholarly Entity Extraction Focused on Machine Learning Models and Datasets. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8166–8176. Singapore: Association for Computational Linguistics.

Pan, H.; Zhang, Q.; Caragea, C.; Dragut, E. C.; and Latecki, L. J. 2024. SciDMT: A Large-Scale Corpus for Detecting Scientific Mentions. In *International Conference on Language Resources and Evaluation*.

Pan, H.; Zhang, Q.; Dragut, E.; Caragea, C.; and Longin, J. L. 2023. DMDD: A Large-Scale Dataset for Dataset Mentions Detection. Available at https://arxiv.org/pdf/2305.11779.pdf.

Pineau, J.; Vincent-Lamarre, P.; Sinha, K.; Lariviere, V.; Beygelzimer, A.; d'Alche Buc, F.; Fox, E.; and Larochelle, H. 2021. Improving Reproducibility in Machine Learning Research(A Report from the NeurIPS 2019 Reproducibility Program). *Journal of Machine Learning Research*, 22(164): 1–20.

Raff, E. 2019. A step toward quantifying independently reproducible machine learning research. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Schindler, D.; Bensmann, F.; Dietze, S.; and Krüger, F. 2021. Somesci-a 5 star open data gold standard knowledge graph of software mentions in scientific articles. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 4574–4583.

sentence-transformers. 2029. multi-qa-mpnet-base-cos-v1 (Hugging Face Model Card). https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-cos-v1. Accessed: 2025-12-16.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Viswanathan, V.; Neubig, G.; and Liu, P. 2021. CitationIE: Leveraging the Citation Graph for Scientific Information Extraction. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 719–731. Online: Association for Computational Linguistics.

Wonsil, J.; Sullivan, J.; Seltzer, M.; and Pocock, A. 2023. Integrated Reproducibility with Self-describing Machine Learning Models. In *Proceedings of the 2023 ACM Conference on Reproducibility and Replicability*, ACM REP '23, 1–14. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701764.

Yan, Z.; Yang, S.; Liu, W.; and Tu, K. 2023. Joint Entity and Relation Extraction with Span Pruning and Hypergraph Neural Networks. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7512–7526. Singapore: Association for Computational Linguistics.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.;

Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Ye, D.; Lin, Y.; Li, P.; and Sun, M. 2022. Packed Levitated Marker for Entity and Relation Extraction. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4904–4917. Dublin, Ireland: Association for Computational Linguistics.

Zhang, Q.; Chen, Z.; Pan, H.; Caragea, C.; Latecki, L. J.; and Dragut, E. 2024. SciER: An Entity and Relation Extraction Dataset for Datasets, Methods, and Tasks in Scientific Documents. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 13083–13100. Miami, Florida, USA: Association for Computational Linguistics.

Zhaohui Yan. 2023. HGERE (GitHub Repository). https://github.com/yanzhh/HGERE. Accessed: 2025-12-16.

Zhong, Z.; and Chen, D. 2021. A Frustratingly Easy Approach for Entity and Relation Extraction. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 50–61. Online: Association for Computational Linguistics.