# Beyond Methods and Datasets Entities: Introducing SH-NER for Hardware and Software Entity Recognition in Scientific Text

**Aftab Anjum**
Kiel University
Germany
afa@informatik.uni-kiel.de

**Nimra Maqbool**
Information Technology University
Lahore, Pakistan
bsce21012@itu.edu.pk

**Ralf Krestel**
Kiel University
Germany
rkr@informatik.uni-kiel.de

## Abstract

Scientific Information Extraction (SciIE) has become essential for organizing and understanding scientific literature, powering tasks such as knowledge graph construction, method recommendation, and automated literature reviews. Although prior SciIE work commonly annotates entities such as tasks, methods, and datasets, it systematically neglects infrastructure-related entities like hardware and software specifications mentioned in publications. This gap limits key applications: knowledge graphs remain incomplete, and recommendation systems cannot effectively filter methods based on hardware compatibility.

To address this gap, we introduce SH-NER, the first large-scale, manually annotated dataset focused on infrastructure-related entities in NLP research. SH-NER comprises 1,128 full-text papers from the ACL Anthology and annotates five entity types: Software, Cloud-Platform, Hardware-Device, Device-Count, and Device-Memory. Our dataset comprises over 9k sample sentences with around 6k annotated entity mentions. To assess the effectiveness of SH-NER, we conducted comprehensive experiments employing state-of-the-art supervised models alongside large language models (LLMs) as baselines. The results show that SH-NER improves scientific information extraction by better capturing infrastructure mentions. You can find the manually annotated dataset at https://github.com/coderhub84/SH-NER.

## 1 Introduction

Scientific Information Extraction (SciIE) (Luan et al., 2017; Groth et al., 2018; Dagdelen et al., 2024) has become a foundational task in enabling machines to understand and organize scientific literature on a scale. With advancements in natural language processing and the availability of annotated corpora, SciIE systems have demonstrated strong performance across core tasks such as scientific named entity recognition (SciNER) (Hong et al., 2020), relation extraction (SciRE) (Zhang et al., 2024a), and citation analysis (Ding et al., 2014). These capabilities power numerous downstream applications, including the construction of scientific knowledge graphs (Dessí et al., 2022), model recommendation systems (Zhao et al., 2024), automated literature review tools (Van Dinter et al., 2021; Orel et al., 2023), scientific question answering (Taffa and Usbeck, 2023; Lehmann et al., 2024), and summarization (Zhang et al., 2024b; Azher et al., 2024).

Although there have been notable advancements in scientific named entity recognition, most existing annotated datasets focus predominantly on entities such as tasks, methods, datasets, evaluation metrics, and citation contexts. For instance, the SciERC (Luan et al., 2017) annotates entities across six scientific types: task, method, metric, material, other-scientific-term, and generic, while SciREX (Jain et al., 2020) captures the contextual grounding of entities by linking them to relevant spans of text across the entire document. Furthermore, some prior work (Te et al., 2022; Mugaanyi et al., 2024) focuses on analyzing citation contexts, whereas S2ORC (Lo et al., 2019) provides a large-scale corpus of full-text scientific papers that are largely unlabeled, containing only minimal annotation. Moreover, none of these resources systematically annotates hardware or software specifications essential for computational reproducibility. This represents a significant gap, as infrastructure-level entities, such as specific software libraries, hardware accelerators, or cloud configurations, are fundamental for understanding experimental setups and ensuring replicability.

The absence of these infrastructure entities has tangible negative impacts: incomplete scientific knowledge graphs and recommendation systems



*Proceedings of Recent Advances in Natural Language Processing*,pages 85–94
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-010

lack the capability to suggest methods compatible with a user's available hardware (e.g., failing to filter out models that require 80GB of GPU memory for users who only have access to 24 GB). Crucially, machine learning models or frameworks struggle to automatically assess replicability without key details, such as software library versions (e.g., PyTorch v1.8 vs. v2.0) or specific hardware models (e.g., RTX 3090 vs. A100). This creates a critical blind spot in the understanding and structuring of computational scientific knowledge.

To address this gap, we introduce the Software Hardware Named Entities Recognition (SH-NER) dataset, the first manually annotated corpus targeting infrastructure-related entities in scientific texts. SH-NER comprises 1,128 full-text papers from the ACL Anthology[1], selected due to NLP's heavy reliance on diverse computational resources and frequent reporting of infrastructure details in experiments. Our dataset focuses on five novel entity types: software entity, cloud platform, device count, device memory, and hardware device. The SH-NER dataset provides granular insights into the computational environments that underpin NLP research.

Annotations were performed by three annotators with a background in computer science, following developed guidelines, achieving an average inter-annotator agreement of $88.63\%$ Fleiss' Kappa score. SH-NER includes 5,287 entity mentions across 3,638 positive entity sentences, along with 5,586 randomly selected negative sentences for comprehensive model training and evaluation (detailed statistics are shown in Table 2). These annotations capture specific details, such as software versions, hardware models, and memory configurations, enabling a clearer understanding and comparison of computational environments.

SH-NER facilitates novel infrastructure-aware SciIE applications, such as:

1. Fine-grained Information Retrieval: Enabling queries like "Find papers fine-tuning BERT-Large using PyTorch and A100 GPUs with 40GB memory."

2. Hardware-Informed Reproducibility Analysis: Automatically flagging papers with underspecified resources (e.g., reporting only "GPU" without specifying model or memory).

3. Enhanced Scientific Knowledge Graphs: Can enrich existing scientific knowledge graphs by integrating hardware and software nodes alongside tasks, methods, and datasets.

4. Enhanced models recommendation: Suggesting models feasible for a user's specific hardware constraints.

We introduce SH-NER, the first manually annotated dataset focused on infrastructure-related entities. To evaluate the effectiveness of SH-NER, we conducted a comprehensive set of experiments with several state-of-the-art supervised baseline models, including BERT-base-uncased, SciBERT, SciDeBERTa-CS, and RoBERTa-Base. In addition to these supervised benchmarks, we also assessed the performance of large language models (LLMs) in a zero-shot setting on the SH-NER dataset, aiming to explore their generalization capabilities in the absence of task-specific fine-tuning.

The remainder of this paper is structured as follows: Section 2 reviews related work on scientific information extraction and benchmark datasets. Section 3 describes the SH-NER dataset, including annotation methodology and statistics. Section 4 details the experimental setup and baseline models. Section 5 presents and analyzes the results. Section 6 concludes with key findings and future directions.

## 2 Related Work

In our research, we have focused on works dealing with the task of scholarly information extraction, particularly in the area of named entity recognition. Although we only looked up methods that are machine learning-based and not rule-based approaches. In general overview (Saier and Färber, 2020) and (Nasar et al., 2018) provide a comprehensive list of information extraction from scientific papers, along with that (Zhang et al., 2024b) and (Xu et al., 2024) provide an in-depth knowledge of using large language models for extracting the information from scientific and non-scientific documents. Multiple datasets have been developed for NER and RE tasks with ground-truth labels; a detailed description and comparison can be found in Table 1.

A significant foundation for research in scientific information extraction (SciIE) has been established through benchmark datasets introduced by the SemEval 2017 and 2018 tasks. The SemEval

[1] ACL Venues

Table 1: Overview of corpora with text scope, entity types, and dataset size.

| Corpus | Scope | Entity Types | #Docs | #Tokens | #Entities |
|---|---|---|---|---|---|
| FTD | titles, abstracts | focus, domain, technique | 426 | 57,182 | 5,382 |
| ACL-RD-TEC | abstracts | lang. resource, resource product, measurement, model, other, tech & method, tool & lib. | 300 | 32,758 | 4,391 |
| TDMS | titles, abstracts, full text | task, dataset, metric, score | 332 | 1,115,987 | 1,384 |
| NGG | titles, abstracts | research problem | 405 | 47,127 | 908 |
| SciERC | abstracts | evaluation metric, generic, material, method, task | 500 | 60,749 | 8,089 |
| Heddes | Sentences | dataset names | 6000 | - | 3,729 |
| GSAP | full text | material,dataset,data Source, method, ml model, model architecture, task | 100 | - | 54,598 |
| SH-NER (Our) | full text | hardware device, device memory, device count, software entity, cloud platform | 1,128 | 572,289 | 5,950 |

2017 dataset (Augenstein et al., 2017) contains 500 paragraphs from scientific papers across computer science, physics, and material science, annotated with three entity types: tasks, methods, and materials. Building upon this, the SemEval 2018 dataset (Gábor et al., 2018) extends the annotation schema to include six relation types, emphasizing intra-sentence relation classification. Along with that, other works have facilitated the development of neural approaches for scientific IE (e.g., (Dagdelen et al., 2024; Liu et al., 2021; Hu et al., 2025; Helwe et al., 2020; Yoon et al., 2019; Gasmi et al., 2024)).

In Information Extraction, researchers annotate different parts of text, including abstracts, sentences, and full texts, etc. (Luan et al., 2018) worked on abstracts and addressed the task of scientific information extraction by jointly modeling entity recognition, relation extraction, and co-reference resolution. They introduce the SCIERC dataset, which contains annotations for all three tasks across 500 scientific abstracts drawn from 12 AI conference proceedings. To tackle the inter-dependencies among these tasks, they propose SCIIE. This unified multi-task framework shares span representations to reduce cascading errors and capture cross-sentence relations via co-reference links. (Heddes et al., 2021) focused on sentence-level annotation for dataset mention detection, introducing a dataset comprising 6,000 manually annotated sentences selected from four major AI conferences based on dataset-related lexical patterns. Other datasets such as FTD (Gupta and Manning, 2011), ACL-RD-TEC (QasemiZadeh and Schumann, 2016), ACL (D'Souza and Auer, 2022), and CL-Titles (D'Souza and Auer, 2021) annotate titles and abstracts to identify various scientific entity types. Further comparative details of these datasets are available in Table 1.

In addition, beyond the sentence level (Jain et al., 2020), SCIREX addresses the challenge of document-level IE by introducing a dataset that supports multiple IE tasks, including salient entity recognition and document-level n-ary relation extraction from scientific articles. To construct the dataset, they combine automated methods with manual annotations, leveraging external scientific knowledge bases to ensure coverage and consistency. Similarly, (Pan et al., 2023) presented DMDD, a large-scale corpus of 31,219 full-text scientific articles, annotated using distant supervision techniques to capture dataset mentions. Meanwhile, (Hou et al., 2021) introduced TDMSci, a dataset targeting task, dataset, and method recognition from 2,000 sentences extracted from NLP publications.

More recently, GSAP (Otto et al., 2023) focuses on enhancing named entity recognition in scientific texts by targeting fine-grained entity types related to machine learning models and datasets. They introduce a manually annotated corpus of 100 full-text scientific publications with 10 entity types, including informal and nested mentions. Their approach emphasizes comprehensive full-text annotation and introduces a baseline model fine-tuned for recognizing both formal and descriptive references to ML-related entities.

Although considerable progress has been made in developing methods and datasets for scientific information extraction (SciIE), existing resources predominantly focus on general entity types or specific aspects, such as citations, methodologies, and

experimental results. In contrast, there is a notable scarcity of datasets explicitly dedicated to the identification and annotation of hardware and software entities within the computational linguistics domain. To address this gap, our work introduces an annotated dataset focused exclusively on hardware and software entities, thereby providing a valuable resource to support and advance research in this relatively underexplored area.

## 3 SH-NER

This section outlines the dataset curation process, covering data collection (3.1), annotation procedures and quality control (3.2–3.3), and annotator agreement, dataset statistics, and comparisons with existing datasets (3.4) to highlight its unique contributions.

### 3.1 Data Acquisition

We collected our dataset from the ACL Anthology[2], a comprehensive repository of publications in the field of computational linguistics. Initially, we retrieved 2,370 publications published between 2020 and 2025. We then filtered these publications using a set of keywords related to hardware specifications, particularly those referencing GPU and CPU usage, resulting in a subset of 1,000 publications. To enhance the diversity of the dataset, we randomly selected an additional 128 publications. In total, we compiled a dataset of 1,128 publications for annotation. These publications cover a range of topics related to machine learning and natural language processing within the domain of computational linguistics. We employed the marker tool[3] to parse the PDF files.

### 3.2 Annotation Tag Set

We defined a set of five annotation tags to systematically label hardware and software-related entities within our dataset.

**Hardware Device**: used to identify physical computing components such as GPUs (e.g., NVIDIA V100, A100) and CPUs (e.g., Intel Xeon).

**Device Memory**: captures mentions of memory size associated with hardware devices, typically expressed in units such as GB (e.g., 32 GB GPU memory).

**Device Count**: annotates numerical references indicating the number of devices used (e.g., 8 GPUs or 2 CPUs).

**Software Entity**: refers to software frameworks, libraries, or tools utilized in the experiments, such as TensorFlow, PyTorch, or Scikit-learn.

**Cloud Platform**: marks references to cloud-based services or infrastructures, such as AWS, Google Cloud, or Azure.

These tags were designed to capture relevant computational and infrastructural details that support reproducibility and transparency in scientific reporting.

### 3.3 Annotation Strategy

The annotation process was carried out by a team of three annotators, all of whom have academic backgrounds in computer science and AI. Before commencing the annotation task, each annotator underwent dedicated training to ensure consistent and reliable annotations for target publications. Among the three annotators, one was a lead annotator who ensured the quality of annotations and ensured that the annotators followed the annotation guidelines. We have developed an annotation guideline used by all annotators throughout the project. This guideline includes instructions for edge cases and particular linguistic cases; we combine the reuse of ACL RD-TEC Guideline[4]. The full annotation guideline can be seen in the data repository link.

A total of 150 publications were randomly selected for joint annotation by all three annotators to evaluate inter-annotator agreement. The remaining documents were evenly divided among the annotators for individual annotation. All annotations were carried out by a predefined set of tags and a detailed annotation guideline. To address specific linguistic phenomena—including determiners, abbreviations, adjectival modifiers, conjunctions, prepositions, and plural forms- we adopted conventions from the ACL RD-TEC Guideline. In particular, determiners such as “a” and “the” were excluded from annotation.

### 3.4 Annotator Agreement & Data statistics

To assess the reliability of our annotation process, we computed the Fleiss’ Kappa score (Davies and Fleiss, 1982) on a subset of 150 publications that were jointly annotated by all annotators. After annotation, we used majority voting to resolve disagreements and select the final labeled samples. The overall inter-annotator agreement for

[2] ACL Venues; NAACL 2025 Long Papers

[3] Marker GitHub Repository

[4] ACL RD-TEC Annotation Guideline (ResearchGate)

these publications was notably high, with an average Fleiss' Kappa score of 88.63%, indicating substantial consensus among annotators. Furthermore, agreement was evaluated separately for each annotation category to ensure consistent labeling across entity types. The Fleiss' Kappa scores for individual labels were as follows: Cloud-Platform 90.74%, Device-Count 83.40%, Hardware-device 95.29%, Software-Entity 81.35%, and Device-Memory 92.37%. These results demonstrate strong and consistent agreement across all entity categories, validating the clarity of the annotation guidelines employed.

The SH-NER dataset comprises 1,128 full-text scholarly documents, containing a total of 5,950 annotated entity mentions across five distinct entity types. Among these, Software Entity (42%) and Hardware Device (27%) are the most frequently occurring, whereas Cloud Platform (4.5%) is the least represented. On average, each document contains approximately 5.64 annotated entities. The dataset comprises 3,638 positive sentences and 5,586 negative sentences; further details can be seen in Table 2.

Notably, the 3,638 positive sentences were not pre-selected for annotation. Instead, annotators reviewed the full text of all 1,128 publications and annotated spans of text where the target entities occurred. Following the annotation process, all sentences containing annotated entities were extracted as positive instances, while a set of non-entity sentences was randomly selected from the same corpus to serve as negative instances.

Furthermore, the number of positive and negative sentences in our dataset is moderately imbalanced for two main reasons. First, this reflects the natural class distribution in real-world full-text scientific documents, where sentences containing entity mentions are typically outnumbered by those without. Second, a larger set of negative examples helps support more robust model training and enhances generalization. To prevent training bias, we maintained a moderate level of imbalance. This design choice enables more realistic learning conditions and contributes to improved performance in practical applications.

## 4 Experimental Setup

### 4.1 Pre-Processing

The SH-NER dataset consists of 1,128 full-text scholarly publications. We used the Marker tool to parse the PDF files into machine-readable text. Before annotation, we performed several preprocessing steps to clean and prepare the data. The parsed output from the Marker tool includes the entire full text of each publication. However, since the majority of our target annotations do not occur within equations or tables, we removed these elements from the parsed text, along with the references section.

In some cases, publications included hardware specifications and hyperparameter settings presented in tabular format. As this study focuses solely on textual content and not on information extraction from tables, such papers were excluded from the corpus.

Additionally, we removed HTML code tags, equations, markdown-style table tags, and non-English characters. Citation spans within the text were also normalized, as the Marker PDF parser often converts them into formats that are inconsistent with the ACL citation style. This normalization ensures that the dataset remains reusable for future research, including the annotation of other entity types.

### 4.2 Dataset splits & Baseline Methods

We split the dataset at the publication level to prevent data leakage, ensuring no sentences from the same publication appear in both training and test sets. The training and test sets comprise 1,008 and 120 publications, respectively, spanning the past five years.

The annotated dataset includes 8,019 training and 1,205 test sentences, with approximately 39% of each containing at least one entity. We used 10% of the training data as a validation set during the supervised model training setup. Moreover, the training and test sets include a total of 5,287 and 663 entities, respectively, averaging 0.66 and 0.55 entities per sentence. Entity-level distribution details are provided in Table 2.

To evaluate the quality and utility of our annotated dataset, we employed four baseline transformer-based models and three large language models (LLMs) across five predefined entity types. The baseline models include BERT-base-uncased (Devlin et al., 2019), SciBERT (Beltagy et al., 2019), SciDeBERTa-CS (Jeong and Kim, 2022), and RoBERTa-Base (Liu et al., 2019), which are widely used for sequence labeling tasks in NLP.

In addition, we assessed the performance of three

| Entity Type | Train | % | Test | % |
|---|---|---|---|---|
| Software-Entity | 2,245 | 42.46 | 281 | 42.38 |
| Hardware-device | 1,478 | 27.95 | 174 | 26.24 |
| Device-Count | 881 | 16.66 | 118 | 17.79 |
| Device-Memory | 507 | 9.6 | 52 | 8.00 |
| Cloud-Platform | 176 | 3.32 | 38 | 5.73 |
| **Total** | 5,287 | 100 | 663 | 100 |

Table 2: Distribution of entity types across training and test sets in the SH-NER dataset, showing both counts and relative percentages.

LLMs: GPT-3.5-turbo, LLaMA-3.3 70B Instruct, and DeepSeek-Chat-v3. This diverse set of model selections enables a thorough assessment of both domain adaptation and generalization capabilities on our annotated data.

### 4.3 Implementation Details

This section outlines the experimental setup for model training and evaluation. Annotated entity spans were converted to token-level labels using the BIO scheme. Four transformer-based models were fine-tuned on our dataset using a learning rate of $2 \times 10^{-5}$, a batch size of 8, and the Adam optimizer with early stopping. Training was conducted for six epochs on a single NVIDIA A100 GPU, with each model requiring approximately 25 to 30 minutes to complete training.

For large language models (LLMs), we evaluated GPT-3.5-Turbo, LLaMA-3-70B-Instruct, and DeepSeek-Chat using the OpenRouter API[5], in both zero- and few-shot learning settings.

**Metrics:** To evaluate the performance of our model on the entity recognition task, we employ three standard evaluation metrics: precision, recall, and F1 score. These metrics are computed under two matching criteria: exact match and partial match, both based on BIO-tagged token sequences. Token-level predictions and ground truth labels are first converted into entity spans by identifying contiguous sequences marked with ”B-“ (begin) and ”I-“ (inside) tags. An exact match is registered when a predicted entity span matches a gold (true) entity in both its span boundaries and entity type. In contrast, a partial match is counted when a predicted and a gold entity share the same type and exhibit any degree of span overlap, even if their boundaries do not align precisely. This dual evaluation strategy enables a comprehensive assessment of the model's ability to both identify entity boundaries and correctly classify entity types.

[5] https://openrouter.ai/models

| Models | Recall | Precision | F1-score |
|---|---|---|---|
| BERT-base-uncased | 81.88 | 87.86 | 84.76 |
| SciBERT | 82.61 | 86.66 | 84.59 |
| SciDeBERTa-CS | 81.23 | 87.28 | 84.14 |
| RoBERTa-Base | 80.68 | 86.40 | 83.44 |
| GPT-3.5-turbo (ZSL) | 53.50 | 26.50 | 35.40 |
| Llama-3.3-70b-instruct (ZSL) | 59.00 | 39.20 | 47.00 |
| Deepseek-chat-v3 (ZSL) | 66.50 | 44.10 | 53.10 |
| GPT-3.5-turbo (FSL) | 57.10 | 28.80 | 37.60 |
| Llama-3.3-70b-instruct (FSL) | 58.00 | 41.20 | 48.15 |
| Deepseek-chat-v3 (FSL) | 65.50 | 48.30 | 55.66 |

Table 3: Exact match performance of supervised models and large language models (LLMs) under zero-shot (ZSL) and few-shot (FSL) settings on the SH-NER dataset.

## 5 Experiment Results

This section presents a comprehensive evaluation of our proposed SH-NER dataset using both fine-tuned supervised models and large language models (LLMs) under zero-shot and few-shot settings. We report performance across exact and partial match criteria to capture both strict boundary accuracy and semantic overlap.

### 5.1 Supervised Baselines

To evaluate the effectiveness of the SH-NER dataset, we conducted experiments using four supervised learning models under two evaluation settings: exact match and partial match. The performance metrics are shown in Tables 3 and 4. The exact match evaluation, which strictly measures entity boundary correctness, shows that all models achieved competitive results. Among them, BERT-base-uncased attained the highest F1-score of 84.76, closely followed by SciBERT and SciDeBERTa-CS. These results suggest that general-purpose pretrained models, when fine-tuned on domain-specific data, can match or exceed the performance of models originally pretrained on scientific text. While SciBERT achieved the highest recall (82.61%), BERT-base-uncased led in precision (87.86%), indicating differing tendencies in error profiles across models.

In the partial match evaluation 4, which tolerates boundary mismatches and better reflects practical

extraction scenarios, all models showed noticeable gains, with F1-scores exceeding. SciDeBERTa-CS achieved the highest score of 90.76%, demonstrating its strength in capturing semantically relevant but loosely defined entity spans. This improvement highlights the value of partial matching for scientific NER tasks, where entity boundaries can be ambiguous. The model's robust performance can be attributed to its contrastive pretraining on scientific corpora, which likely enhances both contextual sensitivity and generalization across varied entity formulations.

Entity-wise performance, as summarized in Table 5, highlights the variability in recognition difficulty across entity categories. Although Software-Entity is the most prevalent type in both training (42.46%) and test (42.38%) sets (see Table 2), it did not yield the highest F1-scores. Instead, Hardware-device and Device-Count entities performed better in terms of exact match F1, likely due to their more distinctive lexical patterns and relatively well-defined contextual usage, despite constituting a smaller share of the data (approximately 28% and 17% in the training set, respectively).

In contrast, Cloud-Platform and Device-Memory categories showed the lowest F1-scores under exact match conditions, dropping to 70.58% in some cases. These lower scores can be attributed to a combination of factors: limited training instances, especially for Cloud-Platform, which makes up only 3.32% of the training data, and higher syntactic and lexical variability. Notably, partial match performance for these challenging categories exhibited considerable improvement, indicating that the model is still able to detect relevant spans even when the exact boundaries are uncertain.

### 5.2 LLMs Baselines

The evaluation of large language models (LLMs) on the SH-NER test set under zero-shot learning (ZSL) and few-shot learning (FSL) settings, as shown in Tables 4 and 3, reveals a consistent performance gap compared to fully supervised transformer models. Under the strict exact match criterion, Deepseek-chat-v3 leads the LLMs, achieving F1-scores of 53.10% (ZSL) and 55.66% (FSL), demonstrating that instruction-tuned LLMs can moderately approximate NER tasks without task-specific fine-tuning. However, these exact match scores remain substantially lower by approximately 25 to 30 points than those of supervised models,

| Models | Recall | Precision | F1-score |
|---|---|---|---|
| BERT-base-uncased | 86.71 | 93.05 | 89.77 |
| SciBERT | 88.38 | 92.72 | 90.50 |
| SciDeBERTa-CS | 87.61 | 94.14 | 90.76 |
| RoBERTa-Base | 86.44 | 92.57 | 89.40 |
| GPT-3.5-turbo (ZSL) | 72.30 | 36.60 | 47.70 |
| Llama-3.3-70b-instruct (ZSL) | 78.20 | 52.80 | 62.30 |
| Deepseek-chat-v3 (ZSL) | 82.20 | 54.10 | 65.30 |
| GPT-3.5-turbo (FSL) | 80.70 | 39.50 | 53.00 |
| Llama-3.3-70b-instruct (FSL) | 76.30 | 58.80 | 66.30 |
| Deepseek-chat-v3 (FSL) | 80.10 | 70.70 | 75.55 |

Table 4: Partial match performance of supervised models and large language models (LLMs) under zero-shot (ZSL) and few-shot (FSL) settings on the SH-NER dataset.

highlighting the difficulty LLMs face in precise entity boundary detection without dedicated training.

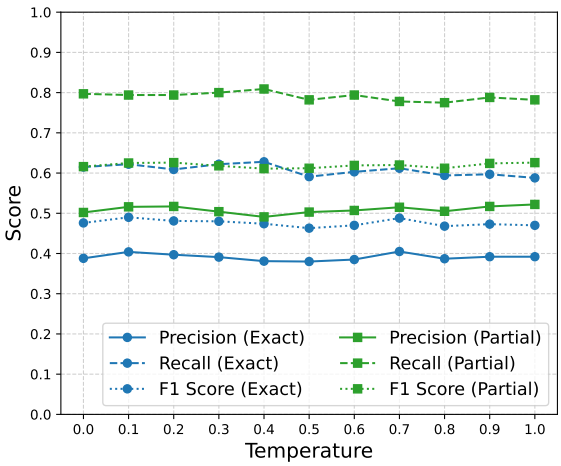


Figure 1: Impact of temperature tuning on LLaMA-3.3-70B-Instruct performance in zero-shot setting.

In contrast, the more lenient partial match evaluation shows significant improvements, with Deepseek-chat-v3 (FSL) attaining an F1-score of 75.55%, substantially closing the gap with supervised approaches. This improvement suggests that while LLMs struggle with exact span extraction, they possess strong semantic recognition abilities when some boundary flexibility is allowed. Overall, these findings indicate that LLMs are becoming increasingly viable for domain-specific NER with limited supervision, although high-precision applications still rely on fine-tuned supervised models. Future research could explore hybrid methods that leverage LLM generalization and supervised

| # Entities | Exact-Match F1 | | | | Partial-Match F1 | | | |
|---|---|---|---|---|---|---|---|---|
| | SciBERT | SciDeBERTa-CS | RoBERTa-Base | BERT-base-uncased | SciBERT | SciDeBERTa-CS | RoBERTa-Base | BERT-base-uncased |
| Hardware device | 87.85 | 83.40 | 84.48 | 89.97 | 96.66 | 95.95 | 95.91 | 96.50 |
| Software Entity | 83.93 | 84.83 | 83.46 | 82.76 | 88.48 | 89.16 | 87.61 | 86.72 |
| Device-Memory | 74.57 | 79.33 | 77.77 | 77.04 | 83.05 | 89.25 | 86.32 | 85.24 |
| Device-Count | 89.83 | 90.98 | 90.29 | 90.98 | 93.22 | 94.42 | 93.67 | 93.56 |
| Cloud Platform | 80.00 | 70.88 | 72.72 | 70.58 | 82.22 | 81.00 | 77.92 | 75.29 |

Table 5: Exact and partial-match F1 scores for five entity types evaluated across four supervised models.

model precision, particularly for complex scientific domains such as hardware and software entity recognition.

We conducted an additional experiment, illustrated in Figure 1, using the LLaMA-3.3-70B-Instruct model due to its faster response rate compared to the Deepseek-chat-v3 model. The objective was to investigate how varying the temperature parameter, which ranged from 0 to 1, influences the model's performance in zero-shot named entity recognition. Temperature is a key hyperparameter controlling the stochasticity and creativity of outputs generated by large language models. We evaluated the effect of different temperature settings on LLaMA's performance using both exact and partial match criteria.

The results demonstrate consistent patterns across temperature variations. For exact match evaluation, recall remained relatively high, but precision was low, resulting in modest F1 scores that peaked at 0.49 with a temperature of 0.1. This indicates a tendency for the model to over-generate entities, favoring recall at the expense of precision. Notably, increasing the temperature beyond 0.1 did not yield further improvements, suggesting a plateau in the tradeoff between response variability and performance. In contrast, partial match metrics showed substantially better performance, with F1 scores consistently above 0.61 and peaking at 0.626 for temperatures of 0.2 and 1.0. These findings suggest that while LLaMA-3.3-70B-Instruct may struggle with exact boundary detection, it effectively captures semantically relevant spans. Overall, lower temperature settings in the range of 0.1 to 0.3 provide the best balance between precision and recall for zero-shot NER using this model.

## 6 Conclusion

In this study, we introduce SH-NER, the first large-scale, manually annotated dataset designed to capture infrastructure-related entities, such as Software libraries/frameworks/tools, Hardware devices, and Cloud-platforms, in the scientific NLP literature. By targeting underrepresented components of infrastructure-related entities, SH-NER fills a substantial gap left by prior SciNER datasets, which primarily focus on tasks, methods, and datasets. Through annotation and experimentation using both supervised learning models and large language models, we demonstrate the usability and effectiveness of SH-NER as a benchmark resource for infrastructure entity recognition. Our findings suggest that while supervised models outperform competitively, large language models show promising results even without fine-tuning.

However, this work also has certain limitations and opens avenues for future exploration. Most notably, the Cloud-Platform entity type is underrepresented in the annotated data, which reflects its broader omission in scientific writing; particularly within the NLP domain. As part of future work, we plan to expand SH-NER to encompass additional scientific disciplines (e.g., computer vision, robotics, and bioinformatics) in order to enhance the diversity and coverage of entity types. Another promising direction for future research is the development of hybrid approaches that combine the generalization capabilities of large language models (LLMs) with the precision of supervised models, particularly in domains such as hardware and software entity recognition. The structured information extracted using such methods could be integrated into open-domain knowledge graphs, thereby enriching them with fine-grained infrastructure-related computational data. This enriched information would be invaluable for downstream applications such as reproducibility assessment, resource-aware literature retrieval, and system-level trend analysis in computational research.

## References


Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics.

Ibrahim Al Azher, Venkata Devesh Reddy Seethi, Akhil Pandey Akella, and Hamed Alhoori. 2024. Limtopic: Llm-based topic modeling and text summarization for analyzing scientific articles limitations. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, pages 1–12.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

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

Mark Davies and Joseph L Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, pages 1047–1051.

Danilo Dessí, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, and Enrico Motta. 2022. Cs-kg: A large-scale knowledge graph of research entities and claims in computer science. In *International Semantic Web Conference*, pages 678–696. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and Chengxiang Zhai. 2014. Content-based citation analysis: The next generation of citation analysis. *Journal of the association for information science and technology*, 65(9):1820–1833.

Jennifer D'Souza and Sören Auer. 2021. Pattern-based acquisition of scientific entities from scholarly article titles. In *International Conference on Asian Digital Libraries*, pages 401–410. Springer.

Jennifer D'Souza and Sören Auer. 2022. Computer science named entity recognition in the open research knowledge graph. In *International Conference on Asian Digital Libraries*, pages 35–45. Springer.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Houssem Gasmi, Jannik Laval, and Abdelaziz Bouras. 2024. Lstm recurrent neural networks for cybersecurity named entity recognition. *arXiv preprint arXiv:2409.10521*.

Paul Groth, Mike Lauruhn, Antony Scerri, and Ron Daniel Jr. 2018. Open information extraction on scientific text: An evaluation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3414–3423, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sonal Gupta and Christopher D Manning. 2011. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proceedings of 5th international joint conference on natural language processing*, pages 1–9.

Jenny Heddes, Pim Meerdink, Miguel Pieters, and Maarten Marx. 2021. The automatic detection of dataset names in scientific articles. *Data*, 6(8).

Chadi Helwe, Ghassan Dib, Mohsen Shamas, and Shady Elbassuoni. 2020. A semi-supervised bert approach for arabic named entity recognition. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 49–57.

Zhi Hong, Roselyne Tchoua, Kyle Chard, and Ian Foster. 2020. Sciner: extracting named entities from scientific literature. In *International Conference on Computational Science*, pages 308–321. Springer.

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.

Haotian Hu, Alex Jie Yang, Sanhong Deng, Dongbo Wang, and Min Song. 2025. Cotel-d3x: A chain-of-thought enhanced large language model for drug–drug interaction triplet extraction. *Expert Systems with Applications*, 273:126953.

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.

Yuna Jeong and Eunhui Kim. 2022. Scideberta: Learning deberta for science technology documents and fine-tuning information extraction tasks. *IEEE Access*, 10:60805–60813.

Jens Lehmann, Antonello Meloni, Enrico Motta, Francesco Osborne, Diego Reforgiato Recupero, Angelo Antonio Salatino, and Sahar Vahdati. 2024. Large language models for scientific question answering: An extensive analysis of the sciqa benchmark. In *European Semantic Web Conference*, pages 199–217. Springer.

Ning Liu, Qian Hu, Huayun Xu, Xing Xu, and Mengxin Chen. 2021. Med-bert: A pretraining framework for medical records named entity recognition. *IEEE Transactions on Industrial Informatics*, 18(8):5600–5608.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. 2019. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific information extraction with semi-supervised neural tagging. *arXiv preprint arXiv:1708.06075*.

Joseph Mugaanyi, Liuying Cai, Sumei Cheng, Caide Lu, and Jing Huang. 2024. Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study. *Journal of Medical Internet Research*, 26:e52935.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2018. Information extraction from scientific articles: a survey. *Scientometrics*, 117(3):1931–1990.

Erol Orel, Iza Ciglenecki, Amaury Thiabaud, Alexander Temerev, Alexandra Calmy, Olivia Keiser, and Aziza Merzouki. 2023. An automated literature review tool (literev) for streamlining and accelerating research using natural language processing and machine learning: Descriptive performance evaluation study. *Journal of medical Internet research*, 25:e39736.

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. *arXiv preprint arXiv:2311.09860*.

Huitong Pan, Qi Zhang, Eduard Dragut, Cornelia Caragea, and Longin Jan Latecki. 2023. Dmdd: A large-scale dataset for dataset mentions detection. *Transactions of the Association for Computational Linguistics*, 11:1132–1146.

Behrang QasemiZadeh and Anne-Kathrin Schumann. 2016. The acl rd-tec 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1862–1868.

Tarek Saier and Michael Färber. 2020. unarxive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *Scientometrics*, 125(3):3085–3108.

Tilahun Abedissa Taffa and Ricardo Usbeck. 2023. Leveraging llms in scholarly knowledge graph question answering. *arXiv preprint arXiv:2311.09841*.

Sonita Te, Amira Barhoumi, Martin Lentschat, Frédérique Bordignon, Cyril Labbé, and François Portet. 2022. Citation context classification: critical vs non-critical. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 49–53.

Raymon Van Dinter, Bedir Tekinerdogan, and Cagatay Catal. 2021. Automation of systematic literature reviews: A systematic literature review. *Information and software technology*, 136:106589.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

Wonjin Yoon, Chan Ho So, Jinhyuk Lee, and Jaewoo Kang. 2019. Collabonet: collaboration of deep neural networks for biomedical named entity recognition. *BMC bioinformatics*, 20:55–65.

Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, and Eduard Dragut. 2024a. Scier: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents. *arXiv preprint arXiv:2410.21155*.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024b. A comprehensive survey of scientific large language models and their applications in scientific discovery. *arXiv preprint arXiv:2406.10833*.

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, et al. 2024. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*.