Bridging the Gap: Instruction-Tuned LLMs for Scientific Named Entity Recognition

Necva Bölücü[1], **Maciej Rybinski**[2], **Stephen Wan**[1]
[1]CSIRO Data61, Sydney, Australia
[2]ITIS, University of Málaga, Málaga, Spain

Abstract

Information extraction (IE) from scientific literature plays an important role in many information-seeking pipelines. Large Language Models (LLMs) have demonstrated strong zero-shot and few-shot performance on IE tasks. However, there are challenges in practical deployment, especially in scenarios that involve sensitive information, such as industrial research or limited budgets. A key question is whether there is a need for a fine-tuned model for optimal domain adaptation (i.e., whether in-domain labelled training data is needed, or zero-shot to few-shot effectiveness is enough). In this paper, we explore this question in the context of IE on scientific literature. We further consider methodological questions, such as alternatives to cloud-based proprietary LLMs (e.g., GPT and Claude) when these are unsuitable due to data privacy, data sensitivity, or cost reasons. This paper outlines empirical results to recommend which locally hosted open-source LLM approach to adopt and illustrates the trade-offs in domain adaptation.

We focus on several instruction-tuning frameworks leveraging IE benchmark datasets to capture task-specific knowledge whilst maintaining model generalisability. We refer to this class of LLM models as *Specialised LLMs* (s-LLMs). We show that instruction-tuned (IE task-adapted) s-LLMs can outperform open-source and proprietary LLMs for entity extraction from scientific documents. Furthermore, this improvement gain is substantial, highlighting the value of the in-domain (continual) fine-tuning.

1 Introduction

Information Extraction (IE) from the scientific literature (e.g., scientific documents, technical reports) is a critical component of scientific information-seeking pipelines (Luan et al., 2018; Nasar et al., 2018; Cai et al., 2025). IE supports tasks such as knowledge-base construction (e.g., BRENDA (Chang et al., 2021) and ChEMBL (Papadatos et al., 2015)), advancing knowledge discovery (Horawalavithana et al., 2022), and supporting predictive modelling (Li et al., 2022). In such pipelines, Named Entity Recognition (NER) is often the initial step used to extract structured output from unstructured text, enabling downstream tasks, such as relation extraction (RE) (Luan et al., 2018) or knowledge-graph construction (Zhang and Soh, 2024). As a result, improving NER accuracy is critical, as errors introduced at this stage can propagate and impact the reliability of the entire pipeline.

As IE pipelines evolve, they are increasingly designed as agentic systems, where multiple specialised models, or agents, collaborate to complete complex tasks (Belcak et al., 2025; Sharma and Mehta, 2025). Within such systems, smaller fine-tuned models play a key role: they can be assigned to specific subtasks, such as NER, RE, or validation, and interact with other agents to balance accuracy, efficiency, and scalability. In this context, NER is not only a technical bottleneck but also a foundational capability for multi-agent scientific systems, motivating the study of models that can be adapted to domain-specific tasks while remaining lightweight and composable.

Recent advances in large language models (LLMs) such as GPT-5[1] and Claude 3.7 Sonnet[2] have improved our ability to extract information from scientific documents. Commercial APIs built in proprietary LLMs offer a strong performance. Using these models becomes problematic, however, in scenarios that involve sensitive data (e.g., biomedical records, confidential industrial research), as privacy cannot be guaranteed. Consequently, many research and industrial settings rely on open-source models as a practical alternative.

Although open-source LLMs provide significant

[1]https://openai.com/gpt-5
[2]https://www.anthropic.com/news/claude-3-7-sonnet

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 56–71
December 23, 2025

flexibility, their zero-shot performance for IE tasks often remains insufficient for practical IE scenarios, as errors propagate to downstream tasks. In-context learning (ICL) enables task and domain adaptation through the inclusion of prototypical examples in the prompt (Li et al., 2023; Ghosh et al., 2024) *without actually performing supervised learning* (no parameter update) called few-shot learning. While ICL markedly improves over zero-shot performance, studies show that it still lags behind state-of-the-art results for IE tasks (Li et al., 2023; Ma et al., 2023; Xu et al., 2024; Wadhwa et al., 2023; Wan et al., 2023; Gao et al., 2023; Jiao et al., 2023; Huang et al., 2024; Wang et al., 2024; Gui et al., 2024b). For the domain of science literature, similar trends have been observed; ICL improves results but does not match supervised fine-tuning models (SLM and LLM) (Xiao et al., 2024; Zhou et al., 2024; Li et al., 2024; Zhang et al., 2025b), and simpler fine-tuned models (e.g., RoBERTa (Liu et al., 2019)) can outperform LLMs using ICL (Jimenez Gutierrez et al., 2022; Bölücü et al., 2023).

To bridge this gap, researchers increasingly turn to instruction-tuned LLMs for IE, which we refer to as *specialised LLMs* (s-LLMs). These models are trained using instruction-tuning on task-specific benchmark datasets (Zhou et al., 2024; Gui et al., 2024b; Wang et al., 2023; Zhang et al., 2025a), where each training instance pairs an instruction, an input text, and a structured output that reflects the benchmark's annotation scheme. Instruction-tuning provides task-level adaptation and enhances zero- and few-shot generalisation, while still enabling local deployment—an essential requirement for domains involving sensitive or proprietary data. Typically built on open-source LLMs such as Llama[3] and Qwen[4], s-LLMs provide cost-effective alternatives to proprietary systems like GPT-4 (Gui et al., 2024b,a; Yuan et al., 2025), making them suitable for applications such as industrial research.

The s-LLMs require a large set of benchmark datasets for instruction-tuning, which is not straightforward and requires substantial computational resources. Therefore, it is not practical to instruction-tune a new model for each conceivable domain for IE. For this reason, in this study, we evaluate the adaptability of *already instruction-tuned* IE-specialised models to scientific domains. Specifically, we focus on three examples of this class of approach: IEPile (Gui et al., 2024b), UniNER (Zhou et al., 2024), and YAYI-UIE (Xiao et al., 2024) (Section 3). These models have been instruction-tuned using a collection of datasets, including scientific datasets (see Table 6), and are designed to generalise across a wide range of IE tasks (e.g., NER, RE, and Event Extraction (EE)) and domains (e.g., social media, biomedical).

Hence, we investigate the following research questions.

- **RQ1:** How well do s-LLMs adapt to the scientific NER task, a subtask of IE, compared to the out-of-the-box (open-source and proprietary) LLMs?
- **RQ2:** What is the additional performance gain of continual (in-domain) tuning of s-LLMs on specific domains compared to their open-source (vanilla) counterparts?

To address the research questions, we evaluate the performance of these models (s-LLMs) and compare them to "out-of-the-box" open-source LLMs[5] (e.g., Llama (Touvron et al., 2023), Baichuan (Yang et al., 2023)), as well as proprietary LLMs (e.g., Claude (Anthropic, 2024), GPT (OpenAI, 2024)). We focus specifically on the case of scientific NER, using four datasets: MeasEval, SciERC, STEM-ECR, and WLPC, each representing a different scientific subdomain or text modality (Section A.4 for an overview of the datasets) in zero-shot, few-shot, and supervised settings. We compare the s-LLMs to baselines under different domain adaptation regimes (zero- and few-shot, continual tuning).

In summary, the **contributions** of this paper include:

- Comparative analysis of instruction-tuned LLMs against their open-source (vanilla) counterparts and proprietary LLMs under different 'learning' regimes (corresponding to different availability of training data).
- Exploratory experiments of models on the NER task to reveal the impact of task-specific instruction-tuning.
- Practical guidelines for researchers aiming to use LLMs for scientific IE.

[3] https://huggingface.co/meta-llama
[4] https://huggingface.co/Qwen

[5] That is, without any further specialisation beyond the foundation model training.

To the best of our knowledge, this is the first extensive evaluation of instruction-tuned LLMs for IE from scientific literature, providing a comprehensive analysis that compares foundation, open-source, and proprietary LLMs and their domain adaptation capabilities across diverse datasets under zero-shot, few-shot and supervised fine-tuning settings.

2 Related Work

LLMs (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023) have already been successfully used in IE (Gao et al., 2023). LLM-based IE methods are divided into In-Context Learning (ICL) and Supervised Fine-tuning (SFT) based approaches. ICL-based models (Jimenez Gutierrez et al., 2022; Li et al., 2023; Wang et al., 2025) rely on prompting with a few labelled examples in addition to instructions, while SFT-based models utilise annotated datasets for fine-tuning LLMs (Zhou et al., 2024; Xiao et al., 2024; Gui et al., 2024b; Li et al., 2024). Research indicates that ICL-based models tend to exhibit relatively inferior effectiveness on IE tasks compared to SFT-based models (Jimenez Gutierrez et al., 2022; Wang et al., 2022; Zhou et al., 2024).

To improve the task and domain adaptability of LLMs, instruction-tuning has become a common technique. This involved fine-tuning LLMs on instruction-based benchmark datasets (a set of datasets specific to a task or domain). Instruction-tuning has been explored across various domains, including Dialogue (Gupta et al., 2022), Intent Classification and Slot Filling (Rosenbaum et al., 2022), Sentiment Analysis (Varia et al., 2023), and Emotion Classification (Liu et al., 2024).

In the context of IE, several studies have advanced instruction-tuning approaches. Zhou et al. (2024) introduce UniNER, which reformulates IE as a Question-Answer (QA) task and instruction-tune Llama using knowledge-distilled datasets from ChatGPT within conversation-style setup, targeting the NER task across diverse domains. Gui et al. (2024b) propose a schema-based instruction-tuning framework for IE (NER, RE and EE) and present IEPile, a bilingual IE instruction benchmark for instruction-tuning. Additionally, Xiao et al. (2024) extend IEPile benchmark by adding more Chinese IE datasets and introduce chat-enhanced instruction tuning that helps gain a fundamental understanding of open-world understanding. Wang et al. (2023) curate the IE INSTRUCTIONS benchmark containing expert-written instructions for diverse IE tasks and apply instruction-tuning for IE tasks. Finally, Lu et al. (2023) focus on *open-world entity profiling*, which is a sub-domain of open-world IE, and construct the INSTRUCTOPEN-WIKI benchmark for the task. They instruction-tune BLOOM to obtain a task-specialised model named PIVOINE.

3 IE-specialised LLMs

Preliminaries Instruction tuning is a supervised fine-tuning (SFT) method in which LLMs are trained on datasets containing human-readable task instructions alongside input-output examples to guide the outputs of LLMs. Each training datapoint, $d = \langle instruction, input, output \rangle$, in the dataset D consists of: (i) an explicit instruction describing the task to be performed; (ii) the corresponding input data; and (iii) the desired output in a defined format.

Unlike standard SFT, which fine-tunes a model on input-output pairs for a specific task without explicit instructions, instruction-tuning conditions the model on natural language task descriptions. This enables better generalisation to unseen domains for the same task (Zhou et al., 2024; Gui et al., 2024b).

Several instruction-tuned LLMs have recently been developed to improve IE performance across diverse domains. We introduce these models (*specialised LLMs for IE*, henceforth 'IE s-LLMs' or simply 's-LLMs') with some discussion of how the approaches vary the basic instruction fine-tuning problem framing.

IEPile (Gui et al., 2024b)[6] proposed a schema-based instruction-tuning, where a schema defines the information to be modelled and extracted, such as entity types, relations, events, etc. This method involves *hard-negative schema construction* and *batched instruction generation*. The schemas are defined as positive (relevant types) and negative (non-relevant types), where negative types can be considered as a kind of "negative" case from a machine learning perspective; the model should not make predictions for this type. To control the complexity of each instruction, the method applies a batching strategy that limits the number of schemas included per instruction using a tunable hyperparameter. The IEPile model training specifically

[6] https://github.com/zjunlp/IEPile

chooses *hard negatives*, labels that are easily confused with positive (i.e., relevant) labels. At inference time, the union set of all schema types across dataset D is presented for prediction.

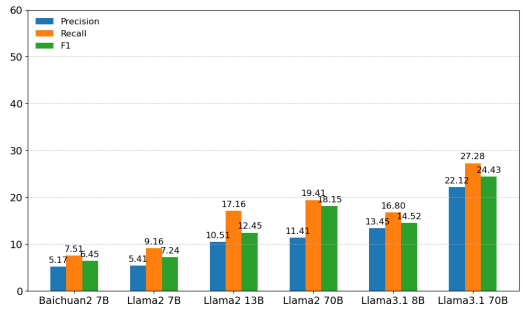
For instruction-tuning, the IEPile benchmark is constructed from a bilingual dataset D that comprises 26 English and 7 Chinese datasets. The dataset spans 3 different tasks: NER, RE, and EE, as exemplified by the datasets ConLL2004 (Carreras and Màrquez, 2004), FabNER (Kumar and Starly, 2022), and BC5CDR (Li et al., 2016), respectively. As a result, the instruction will differ for these datasets, with content specific to each of the task descriptions.

UniNER (Zhou et al., 2024)[7] is a framework that uses ChatGPT for knowledge distillation to generate instruction-tuning data for the NER task. It uses broad-coverage, unlabeled web text and distils this information into an instruction-tuned model built on an open-source LLM (LLaMA), resulting in the UniversalNER models.

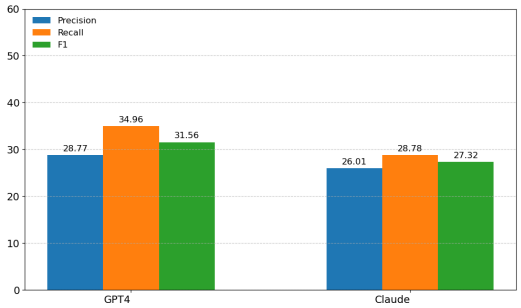
Unlike traditional NER, which frames the task as entity detection, UniNER reformulates it as a question-answering (QA) task. The model input is a question about what entity is present in the accompanying text (e.g., What describes t_1 in the text?), and the output is the corresponding entity span. These QA pairs are generated using GPT, which is prompted to answer such questions based on given texts. The responses are collected as "conversation" transcripts and subsequently segmented into QA tuples t, forming a training dataset for instruction tuning. In data construction, they apply *negative sampling* where non-relevant entity types are included in the dataset. This process creates a distilled dataset suitable for fine-tuning LLaMA-2 (Touvron et al., 2023), resulting in instruction-tuned models that generalise well across domains.

Additionally, the authors introduce a benchmark dataset D consisting of 43 datasets from a wide range of domains, including biomedicine, law, and finance, to evaluate models.

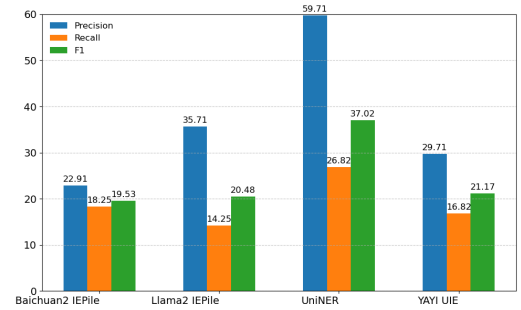
YAYI-UIE (Xiao et al., 2024)[8] is an instruction-tuning framework that consists of two steps: (i) instruction-tuning for chat, where an open-source dialogue data with instructions and a self-constructed corpus is used to train a chat-enhanced language model to gain a fundamental understanding of open-world language and enhance Chinese language capabilities. A key step in the chat-based training is to filter low-quality samples, such as meaningless, incomplete, sensitive, or duplicate samples; (ii) instruction-tuning for IE, where the chat-based model is used to tune for IE tasks with a benchmark dataset. The benchmark D includes a combined dataset of 16 Chinese IE datasets and the InstructUIE benchmark (Wang et al., 2023) for IE instruction-tuning, spanning data from diverse sources such as finance, politics, and security.



(a) Open-source LLMs



(b) Proprietary LLMs



(c) s-LLMs

Figure 1: Zero-shot performance comparison of open-source, proprietary, and IE s-LLMs on the SciERC dataset.

Statistical details of scientific datasets used in instruction-tuning of IE (s-LLMs) are given in Appendix A.4 in Table 6.

[7] github.com/universal-ner/universal-ner
[8] huggingface.co/wenge-research/yayi-uie

Method	SciERC		Stem-ECR		MeasEval		WLPC	
	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot
Proprietary LLMs								
GPT-4	31.56	41.12	21.39	**35.59**	15.67	24.47	<u>52.95</u>	**59.16**
Claude 3.5 Sonnet	27.32	34.19	21.70	34.06	14.70	22.47	36.21	41.05
Open-source LLMs								
Baichuan2	6.45	11.56	8.18	14.12	12.48	16.18	6.75	14.10
Llama2 7B	7.24	13.20	9.47	15.01	11.18	18.14	10.67	19.78
Llama2 13B	12.45	22.38	10.52	19.42	11.36	19.21	12.44	21.45
Llama2 70B	18.12	26.89	12.89	20.17	12.45	20.56	15.74	22.49
Llama3.1 8B	14.52	24.80	10.14	20.17	11.72	20.10	9.32	19.17
Llama3.1 70B	24.43	29.45	14.73	21.45	12.14	21.49	18.42	25.19
IE s-LLMs								
Baichuan2-IEPile	19.53	32.15	15.22	23.10	16.78	26.32	28.79	35.17
Llama2-IEPile	20.48	18.49	18.49	25.42	24.18	29.12	30.40	36.56
UniNER-7B	<u>37.01</u>	**46.43**	17.26	24.37	11.29	19.47	30.13	35.49
YAYI-UIE	22.55	35.19	<u>25.89</u>	32.17	<u>24.62</u>	**30.15**	33.13	37.10

Table 1: Zero-shot and few-shot (1-shot) performance on NER datasets (of note, SciERC is in-domain for the three s-LLMs (IEPile, UniNER and YAYI-UIE); the other datasets are out-of-domain). The best zero-shot results for each dataset are <u>underlined</u>, and the best few-shot results for each dataset are **boldfaced**.

4 Results and Analyses

We design our experiments to evaluate the performance of s-LLMs on the scientific NER task[9]. Our goal is threefold: (1) evaluation of zero-shot and few-shot (1-shot) capabilities of s-LLMs against their open-source (vanilla) counterparts and proprietary LLMs; (2) comparison of continual fine-tuning (in-domain) of s-LLMs' performance against their open-source counterparts' fine-tuning; and (3) exploration of the generalisability of in-domain adapted models to a specific dataset to other scientific datasets. Finally, we present experimental results on the general domain to compare them with findings from the scientific domain.

4.1 Experiments in Zero-shot and Few-shot Settings for Scientific Domain

Our first experiment focuses on examining whether the entity extraction capability learned by s-LLMs is transferable across scientific domain datasets under zero-shot and few-shot settings (**RQ1**). Table 1 reports performance across datasets. Of these, only **SciERC** was used during the instruction-tuning of the s-LLMs and is thus considered the *in-domain* (seen in training) dataset (see Table 6)[10]. The remaining datasets (STEM-ECR, MeasEval and WLPC) are *out-of-domain* (unseen in training), representing unseen entity type sets (covariate shift) and datasets.

Table 1 enables direct comparison across the prior work for the first time. Here, our results include results for open-source and proprietary LLMs that are state-of-the-art at the time of writing.

To begin with, as expected, we note that proprietary LLMs (GPT-4 and Claude 3.5 Sonnet) stand out as strong baselines across the board. Their performance is particularly impressive given the presumed absence of task-specific fine-tuning. Their effectiveness is the best/second-highest performance on most datasets. This demonstrates their ability to generalise across interpretable entity types (e.g., SciERC: *Material, Method, Metric, ⋯*; STEM-ECR: *Data, Material, ⋯*; WLPC: *Ph, Size, Action, ⋯*). However, our aim in this paper is to explore the best methods to obtain alternatives to these cloud-based models, which may be locally hosted by an organisation (particularly if they are responsible for sensitive data). We thus turn our focus to open-source LLMs.

In general, we find that zero-shot inference from IE s-LLMs is better than using open-source LLMs without any task specialisation. For SciERC, UniNER-7B (based on Llama2-7B) achieves a higher F_1 score than both open-source and proprietary LLMs. This demonstrates the benefit of task-specific instruction-tuning. Note that the SciERC dataset is used for the NER task in the UniNER model, whereas it is used as the RE dataset for the IEPile (based on Llama2-13B & Baichuan2-13B)

[9]Experimental settings are given in Appendix A.

[10]Results on it are not strictly zero-shot.

and YAYI-UIE (based on Baichuan2-13B) models (see Table 6). Indeed, the s-LLMs, UniNER-7B and YAYI-UIE, generally outperform the proprietary models for all the datasets except WLPC (which includes text from technical documentation instead of scientific publications), which is particularly interesting given the generally smaller parameter size of s-LLMs compared to GPT-4 and Claude 3.5 Sonnet. However, we note that the margin only has a maximum difference of approximately 9 F_1 points in the case of MeasEval (YAYI-UIE vs GPT-4). In a few-shot setting (1-shot), all models (open-source, proprietary and s-LLMs) benefit from ICL examples, leading to performance gains over zero-shot baselines. Excluding proprietary LLMs, the trend remains consistent. s-LLMs outperform their open-source(vanilla) counterparts.

To understand why s-LLMs exhibit performance gains, we analyse the precision and recall metrics for the models (open-source, proprietary and s-LLMs), presented in Figure 1. This figure presents a comparative analysis of zero-shot performance on the SciERC dataset. Notably, the s-LLMs lead to increased precision, at the expense of recall. In the case of the UniNER approach, the precision gains strongly outweigh any drop in recall. This indicates that targeted training on IE tasks enhances the models' ability to identify relevant entities with greater accuracy. Additionally, these models tend to be relatively conservative and precise in their positive predictions, though they may miss some relevant instances.

In conclusion, while s-LLMs benefit from fine-tuning, they still face generalisation challenges in scientific domains (i.e., the low recall). Moreover, although the s-LLMs are competitive against proprietary LLMs, the performance gap remains narrow in some cases, underscoring the need for further advancements in training and fine-tuning strategies to improve robustness. As a result, we turn our attention to the continued fine-tuning of the IE capability of both open-source LLMs and IE s-LLMs for supervised domain adaptation.

4.2 On the Benefits of Continual In-domain Fine-tuning for Scientific Writing

The results from the previous section show that IE s-LLMs remain competitive against proprietary LLMs under zero-shot and few-shot settings. However, despite their strengths, a performance gap remains compared to SFT models in scientific domain datasets, indicating that there is still room for further improvement.

In this section, we ask: does continual in-domain tuning on the *target* dataset lead to additional performance gains, or do IE s-LLMs already reach peak performance on scientific datasets through their general instruction-tuning? (**RQ2**) In the context of our motivation in Section 1, one might consider how further fine-tuning of a local model on a sensitive or private dataset might improve results.

Following prior work (Zhou et al., 2024; Gui et al., 2024b), we refer to this addition as continual in-domain fine-tuning, a next step after instruction-tuning that further adapts the model to a specific dataset and denote this in our results tables as SFT (for supervised fine-tuning).

To explore the impact of continual in-domain fine-tuning, we fine-tune both open-source (vanilla) LLMs and IE s-LLMs using the training sets of the scientific datasets (Appendix A.4). Table 2 presents the results of the SFT regime compared to zero-shot performance, alongside GPT-4 (zero-shot) and BERT-base (fine-tuned) as baselines; BERT-base represents the task-specific supervised models commonly used across studies in the literature (Xiao et al., 2024; Zhou et al., 2024; Gui et al., 2024b). The table shows that all SFT models improve significantly on all datasets compared to their untuned counterparts. Notably, they outperform GPT-4 in zero-shot settings by a considerable margin. For example, for the STEM-ECR dataset, the difference is over 55 F_1 points, demonstrating clearly that fine-tuning is still a preferred approach in the presence of annotated training data.

The results demonstrate that in-domain fine-tuning on a specific dataset helps, whether this is the original open-source LLMs or the s-LLMs. However, performance gains from in-domain fine-tuning are greater when starting with IE s-LLMs, indicating learning from the multiple datasets used in the s-LLMs training is transferable, demonstrating the benefits of instruction tuning and subsequent in-domain optimisation. Among the in-domain fine-tuned models, the YAYI-UIE model achieves the highest Micro F_1 score among the SFT models across all datasets, showing its strong performance in NER. This reflects its ability to handle diverse scientific NER tasks, possibly related to its larger benchmark datasets covering a wide range of IE tasks and domains in instruction tuning. YAYI-UIE differs from other methods (UniNER and IEPIle) in that dialogue data is used to perform general instruction tuning to train a chat-enhanced

	SciERC		STEM-ECR		MeasEval		WLPC	
Model	**Zero-shot**	**SFT**	**Zero-shot**	**SFT**	**Zero-shot**	**SFT**	**Zero-shot**	**SFT**
Open-source LLMs								
Baichuan2	6.45	52.18	8.18	51.08	12.48	48.47	6.75	35.23
Llama2 7B	7.24	53.14	9.47	50.98	11.18	52.78	10.67	39.56
Llama2 13B	12.45	55.45	10.52	57.14	11.36	54.10	12.44	42.21
Llama2 70B	15.12	56.48	12.89	59.24	12.45	53.18	15.74	45.40
Llama3.1 8B	14.52	56.20	10.14	56.74	11.72	54.10	9.32	43.18
Llama3.1 70B	24.43	55.26	14.73	58.31	12.14	52.85	18.42	46.12
LLMs optimised for IE tasks								
Baichuan2-IEPile	19.53	73.18	15.22	75.12	16.78	59.14	28.79	60.19
Llama2-IEPile	20.48	76.08	18.49	78.17	24.18	64.10	30.40	62.58
UniNER-7B	37.01	78.41	17.26	79.02	11.29	66.18	30.13	60.45
YAYI-UIE	21.17	**80.47**	25.89	**82.52**	24.62	**69.71**	33.13	**64.17**
BERT-base	-	62.81 ±0.85	-	68.17 ±0.76	-	55.43 ±1.15	-	39.52 ±0.52
GPT-4	31.56	-	21.39	-	15.67	-	52.95	-

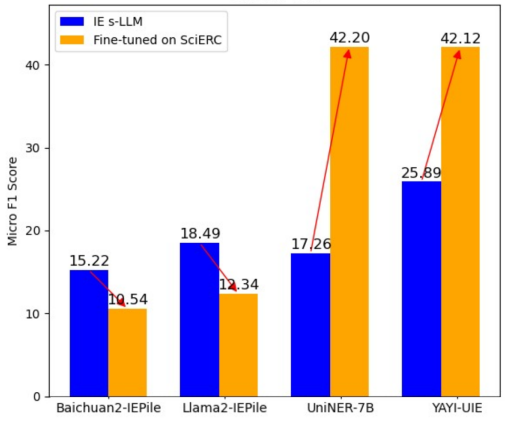
Table 2: Strict Micro F_1 on NER datasets for zero-shot and SFT settings. The best zero-shot results for each dataset are underlined, and the best SFT results for each dataset are **boldfaced**.

model using a dialogue corpus in both English and Chinese instead of using an instruction model.

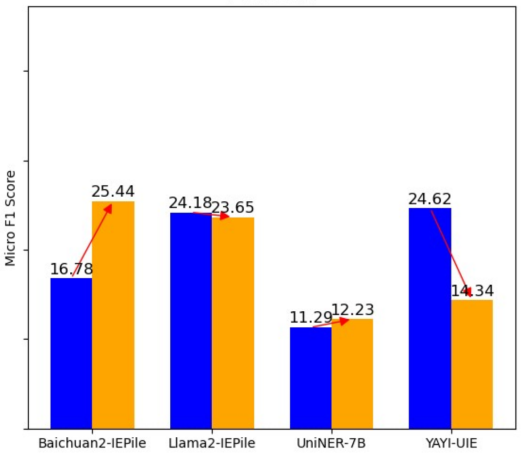
Of note, SFT appears somewhat ineffective for base open-source LLMs. Specifically, the BERT baseline yielded higher effectiveness on most datasets. While IE s-LLMs achieved the best performance among all SFT models, this comes at a cost. These models require extensive data resources for training (YAYI-UIE: *49 datasets*, IEPile: *33 datasets*, and UniNER: *43 datasets*) and significant computational resources for instruction-tuning and supervised fine-tuning compared to fine-tuning PLMs for the NER task. The model complexity alone can limit their accessibility and scalability for researchers or practitioners with resource constraints.

This highlights a fundamental trade-off between effectiveness and efficiency: IE s-LLMs deliver state-of-the-art performance but with higher training and inference cost, while smaller models like BERT offer a practical balance of accuracy and affordability.

In summary, continual fine-tuning remains critical for achieving optimal performance in scientific IE. When paired with general instruction tuning, this two-stage process supports both generalisability and domain specialisation (dataset adaptation), enabling robust and adaptable solutions for real-world applications.



(a) STEM-ECR



(b) MeasEval

Figure 2: Strict Micro F_1 on NER datasets obtained from IE s-LLMs and fine-tuned on the SciERC dataset.

4.3 Generalisability of Fine-tuned s-LLMs

To assess whether the continual in-domain fine-tuning also leads to generalisable models (to other scientific datasets), we take the IE s-LLM models obtained by continual fine-tuning on the SciERC dataset (`X-SciERC`) and use these models for zero-shot inference on the MeasEval and STEM-

ECR datasets. We choose the SciERC dataset because there is an entity type overlap with the STEM-ECR dataset ('Material', 'Method'), but not with the MeasEval dataset. The results are presented in Figure 2.

The findings indicate that IEPile models fine-tuned in-domain on the SciERC dataset (`X-IEPile-SciERC`) exhibit lower performance on the STEM-ECR dataset, while the UniNER and YAYI-UIE models demonstrate improved performance. The reason behind this might be the knowledge distillation used in the instruction-tuning of UniNER and the larger benchmark used in the tuning of YAYI-UIE and UniNER models. For the MeasEval dataset, the `UniNER-7B-SciERC` model provides a slight improvement, and `Baichuan2-IEPile-SciERC` outperforms the zero-shot `Baichuan2-IEPile`. In contrast, the continually trained YAYI-UIE model yields a performance drop.

From these results, we conclude that the general applicability of the model depends on how close the out-of-domain data is to the data used for continual training. As the SciERC and STEM-ECR entity types share some overlap (being about general concepts relating to the scientific method), we observe better cross-domain effectiveness in UniNER and YAYI-UIE models. In contrast, for the MeasEval dataset, given its particular focus on quantitative measurements, we see no meaningful improvements stemming from out-of-domain training, and, in one case (the YAYI-UIE model), we actually observe a marked performance drop.

4.4 General Domain Evaluation

To assess the generalisability of our findings to domains beyond scientific information extraction, we evaluated s-LLMs using the CrossNER (Liu et al., 2021) and CoNLL2003 (Tjong Kim Sang and De Meulder, 2003) datasets (statistical details are provided in Table 5). ConLL2003 is used for in-domain and CrossNER is used for out-of-domain, as outlined by Wang et al. (2023); Zhou et al. (2024).

As shown in Table 3, s-LLMs outperform their open-source (vanilla) counterparts on the in-domain dataset (CoNLL2003). However, GPT-4 (a proprietary LLM) outperforms these models on the CrossNER dataset. This performance gap may be related to two possible factors: (i) the model's undisclosed pretraining data, which may include broader coverage of domains or overlap with similar data; and (ii) a similar trend observed in scientific domain evaluations (Section 4.1), where s-LLMs struggle with generalisation to unseen datasets.

These findings show a key limitation of s-LLMs: instruction-tuning improves performance within domains present in the instruction-tuning data; however, it does not guarantee robustness to domain shifts. In contrast, large-scale proprietary LLMs like GPT-4 benefit from diverse pretraining data or emergent generalisation capabilities (although these are difficult to verify given the lack of transparency around the training data and regime).

Model	CoNLL2003	AI	Literature	Music	Politics	Science
Proprietary LLMs						
GPT-4	68.68	**61.95**	52.32	**70.79**	63.99	62.66
Claude 3.5 Sonnet	55.10	32.78	30.18	43.52	45.37	47.12
Open-source LLMs						
Baichuan2	20.50	4.17	12.14	16.89	20.47	8.52
Llama2 7B	17.06	5.19	13.87	17.42	11.96	9.24
Llama2 13B	33.47	13.92	28.92	33.96	36.97	23.85
Llama2 70B	43.39	39.10	40.67	49.30	53.49	39.50
Llama3.1 8B	62.48	40.12	42.17	48.82	30.15	45.12
Llama3.1 70B	70.47	51.42	56.08	64.02	38.34	52.49
IE s-LLMs						
Baichuan2-IEPile	70.41	56.12	50.52	59.18	53.17	55.10
Llama2-IEPile	72.40	53.47	62.15	58.72	55.67	57.68
UniNER-7B	**81.14**	60.25	**62.98**	66.35	**65.30**	**69.23**
YAYI-UIE	78.18	51.60	43.38	61.46	47.43	48.45

Table 3: Zero-shot performance on general domain NER datasets (CoNLL2003 is in-domain; CrossNER is out-of-domain). The best results are **boldfaced**.

4.5 Practical Recommendations

Based on our evaluation of s-LLMs compared to proprietary and open-source LLMs for the scientific domain, we make the following recommendations for practitioners, especially those working in privacy-sensitive or resource-constrained environments in the domain of scientific literature information extraction:

1. **Domain adaptation as a solution for local (open-source) models.** For open-source LLMs, task adaptation (instruction-tuning) is required to enhance the task-specific zero- and few-shot generalisation capabilities of LLMs; i.e., open-source models without it perform poorly, perhaps too poorly for prototyping.
2. **s-LLMs as a starting point for dataset adaptation.** For in-domain adaptation in the scientific domain, starting with an s-LLM that has already adapted to the task yields stronger performance. The prior multi-task training often provides a useful foundation that can

be transferred across domains. On the flipside, direct instruction tuning of a base open-source model provides limited value (or requires much more training data, see below).

3. **YAYI-UIE demonstrates the best overall performance and generalisation across s-LLMs.** Among s-LLMs, YAYI-UIE achieves the highest and most consistent results after continual in-domain fine-tuning, making it a strong choice for scientific IE applications.
4. **Task adaptation with a larger benchmark.** Gathering in-domain training data and using it for instruction-tuning is still the most effective way of task adaptation. For LLMs, instruction-tuning for task adaptation appears to require a prior step (instruction tuning), as direct in-domain fine-tuning vanilla open-source LLMs appears to yield subpar results.
5. **Smaller PLMs remain viable cost-effective alternatives.** Although s-LLMs offer improved performance, smaller PLMs like BERT can still provide competitive results, if in-domain training data can be sourced. Their lower computational demands make them practical options for projects with limited resources.

5 Conclusion

In this paper, we investigate instruction-tuned IE specialised LLMs (s-LLMs), specifically focusing on their performance in scientific entity extraction compared to open-source and proprietary LLMs. The experimental results show that s-LLMs perform better than their open-source (vanilla) counterparts, showing that instruction-tuning benefits in task-adaptation. However, s-LLMs still face a generalisation problem in the scientific domain. Continual in-domain fine-tuning of IE s-LLMs leads to the best results, particularly for specific scientific datasets of interest. In our experiments, these models outperformed proprietary ones by up to an order of magnitude, achieving over 55 F_1 points in zero-shot and 20 F_1 points in few-shot settings.

We also observe that models like YAYI-UIE perform well across a variety of datasets, highlighting their adaptability to unseen datasets in zero-shot and few-shot settings. However, the choice of s-LLM and its suitability for a given dataset remains a hyperparameter defined in the study. Despite the success of s-LLMs, PLMs (BERT) continue to offer competitive and cost-effective alternatives for NER, particularly when in-domain train data is available, often outperforming open-source LLMS in-domain tuned directly for specific tasks.

This work highlights the strengths and weaknesses of s-LLMs in scientific NER and provides a comparative analysis across zero-shot, few-shot and fine-tuned settings. However, our study is limited in scope: we focused exclusively on sentence-level NER within the scientific domain and relied on publicly available s-LLMs without modifications. As such, the performance and limitations of these models inherently constrain our findings. Additionally, due to resource limitations, we did not evaluate large proprietary LLMs such as GPT-4 or Claude under fine-tuned conditions. We also did not explore the problem of catastrophic forgetting in s-LLMs, which is important to understand how well these models retain knowledge and problem-solving skills learned from previous tasks.

Future work will extend this evaluation to other IE tasks such as relation and event extraction, and investigate how combining the strengths of different s-LLMs (e.g., UniNER's strong zero-shot performance vs. YAYI-UIE's fine-tuning responsiveness) can lead to more robust pipelines. Expanding the diversity and number of datasets may also help in identifying better general-purpose starting points for scientific information extraction.

Limitations

Our study is centred exclusively on the sentence-level Named Entity Recognition (NER) task. Specifically, we concentrate on the scientific domain, which may require further exploration to apply our findings to other domains. Additionally, due to resource constraints, we were unable to fine-tune large language models with more parameters (e.g., GPT-4, Claude). We use the IE s-LLMs provided by the papers. The limitations derived from these models are also limitations of our study.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Online. Accessed: 2024-08-13.

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. 2025. Small Lan-

guage Models are the Future of Agentic AI. *Preprint*, arXiv:2506.02153.

Necva Bölücü, Maciej Rybinski, and Stephen Wan. 2023. impact of sample selection on in-context learning for entity extraction from scientific writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5090–5107, Singapore. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. 33:1877–1901.

Ruichu Cai, Junhao Lu, Zhongjie Chen, Boyan Xu, and Zhifeng Hao. 2025. Handling Missing Entities in Zero-Shot Named Entity Recognition: Integrated Recall and Retrieval Augmentation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10790–10802, Albuquerque, New Mexico. Association for Computational Linguistics.

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 89–97, Boston, Massachusetts, USA. Association for Computational Linguistics.

Antje Chang, Lisa Jeske, Sandra Ulbrich, Julia Hofmann, Julia Koblitz, Ida Schomburg, Meina Neumann-Schaal, Dieter Jahn, and Dietmar Schomburg. 2021. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic acids research*, 49(D1):D498–D508.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Jennifer D'Souza, Anett Hoppe, Arthur Brack, Mohmad Yaser Jaradeh, Sören Auer, and Ralph Ewerth. 2020. The STEM-ECR Dataset: Grounding Scientific Entity References in STEM Scholarly Content to Authoritative Encyclopedic and Lexicographic Sources. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2192–2203, Marseille, France. European Language Resources Association.

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange. 2020. The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1255–1268, Online. Association for Computational Linguistics.

Jun Gao, Huan Zhao, Yice Zhang, Wei Wang, Changlong Yu, and Ruifeng Xu. 2023. Benchmarking large language models with augmented instructions for fine-grained information extraction. *arXiv preprint arXiv:2310.05092*.

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

Felix Grezes, Sergi Blanco-Cuaresma, Thomas Allen, and Tirthankar Ghosal. 2022. Overview of the First Shared Task on Detecting Entities in the Astrophysics Literature (DEAL). In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 1–7, Online. Association for Computational Linguistics.

Honghao Gui, Shuofei Qiao, Jintian Zhang, Hongbin Ye, Mengshu Sun, Lei Liang, Jeff Z Pan, Huajun Chen, and Ningyu Zhang. 2024a. Instructie: A bilingual instruction-based information extraction dataset. In *International Semantic Web Conference*, pages 59–79. Springer.

Honghao Gui, Lin Yuan, Hongbin Ye, Ningyu Zhang, Mengshu Sun, Lei Liang, and Huajun Chen. 2024b. IEPile: Unearthing Large Scale Schema-Conditioned Information Extraction Corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 127–146, Bangkok, Thailand. Association for Computational Linguistics.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey Bigham. 2022. InstructDial: Improving Zero and Few-shot Generalization in Dialogue through Instruction Tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 505–525, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Corey Harper, Jessica Cox, Curt Kohler, Antony Scerri, Ron Daniel Jr., and Paul Groth. 2021. SemEval-2021 Task 8: MeasEval – Extracting Counts and Measurements and their Related Contexts. In *Proceedings of*

the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 306–316, Online. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

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

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. 2024. TextEE: Benchmark, Reevaluation, Reflections, and Future Challenges in Event Extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12804–12825, Bangkok, Thailand. Association for Computational Linguistics.

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.

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

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chaitanya Kulkarni, Wei Xu, Alan Ritter, and Raghu Machiraju. 2018. An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 97–106, New Orleans, Louisiana. Association for Computational Linguistics.

Aman Kumar and Binil Starly. 2022. "FabNER": information extraction from manufacturing process science domain literature using named entity recognition. *Journal of Intelligent Manufacturing*, 33(8):2393–2407.

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

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

Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. 2023. CodeIE: Large Code Generation Models are Better Few-Shot Information Extractors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Toronto, Canada. Association for Computational Linguistics.

Zixuan Li, Yutao Zeng, Yuxin Zuo, Weicheng Ren, Wenxuan Liu, Miao Su, Yucan Guo, Yantao Liu, Lixiang Lixiang, Zhilei Hu, Long Bai, Wei Li, Yidan Liu, Pan Yang, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. 2024. KnowCoder: Coding Structured Knowledge into LLMs for Universal Information Extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8758–8779, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A Joint Neural Model for Information Extraction with Global Features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

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*.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. EmoLLMs: A Series

of Emotional Large Language Models and Annotation Tools for Comprehensive Affective Analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 5487–5496, New York, NY, USA. Association for Computing Machinery.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. CrossNER: Evaluating Cross-Domain Named Entity Recognition. volume 35, pages 13452–13460.

Keming Lu, Xiaoman Pan, Kaiqiang Song, Hongming Zhang, Dong Yu, and Jianshu Chen. 2023. PIVOINE: Instruction Tuning for Open-world Entity Profiling. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15108–15127, Singapore. Association for Computational Linguistics.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified Structure Generation for Universal Information Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

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

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10572–10601, Singapore. Association for Computational Linguistics.

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

OpenAI. 2024. Hello gpt-4o. Accessed: 2024-08-13.

Yixin Ou, Ningyu Zhang, Shengyu Mao, Runnan Fang, Yinuo Jiang, Ziwen Xu, Xiaolong Weng, Lei Li, Shuofei Qiao, and Huajun Chen. 2023. EasyInstruct: An Easy-to-use Framework to Instruct Large Language Models. `https://github.com/zjunlp/EasyInstruct`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. 35:27730–27744.

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

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. 2022. LINGUIST: Language Model Instruction Tuning to Generate Annotated Utterances for Intent Classification and Slot Tagging. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 218–241, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

David Schindler, Felix Bensmann, Stefan Dietze, and Frank Krüger. 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*, CIKM '21, page 4574–4583, New York, NY, USA. Association for Computing Machinery.

Raghav Sharma and Manan Mehta. 2025. Small Language Models for Agentic Systems: A Survey of Architectures, Capabilities, and Deployment Trade offs. *Preprint*, arXiv:2510.03847.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Siddharth Varia, Shuai Wang, Kishaloy Halder, Robert Vacareanu, Miguel Ballesteros, Yassine Benajiba, Neha Anna John, Rishita Anubhai, Smaranda Muresan, and Dan Roth. 2023. Instruction Tuning for Few-Shot Aspect-Based Sentiment Analysis. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 19–27, Toronto, Canada. Association for Computational Linguistics.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting Relation Extraction in the era of Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context Learning for Relation Extraction using Large Language Models. In *Proceedings of the*

2023 Conference on Empirical Methods in Natural Language Processing, pages 3534–3547, Singapore. Association for Computational Linguistics.

Jiaqi Wang, Yuying Chang, Zhong Li, Ning An, Qi Ma, Lei Hei, Haibo Luo, Yifei Lu, and Feiliang Ren. 2024. Techgpt-2.0: A large language model project to solve the task of knowledge graph construction. *arXiv preprint arXiv:2401.04507*.

Liwen Wang, Rumei Li, Yang Yan, Yuanmeng Yan, Sirui Wang, Wei Wu, and Weiran Xu. 2022. Instructionner: A multi-task instruction-based generative framework for few-shot ner. *arXiv preprint arXiv:2203.03903*.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named Entity Recognition via Large Language Models. pages 4257–4275.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023. InstructUIE: multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xinglin Xiao, Yijie Wang, Nan Xu, Yuqi Wang, Hanxuan Yang, Minzheng Wang, Yin Luo, Lei Wang, Wenji Mao, and Daniel Zeng. 2024. Yayi-uie: A chat-enhanced instruction tuning framework for universal information extraction.

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.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Lin Yuan, Jun Xu, Honghao Gui, Mengshu Sun, Zhiqiang Zhang, Lei Liang, and Jun Zhou. 2025. Improving natural language understanding for llms via large-scale instruction synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25787–25795.

Bowen Zhang and Harold Soh. 2024. Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9820–9836, Miami, Florida, USA. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2025a. Instruction Tuning for Large Language Models: A Survey.

Zikang Zhang, Wangjie You, Tianci Wu, Xinrui Wang, Juntao Li, and Min Zhang. 2025b. A Survey of Generative Information Extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4840–4870, Abu Dhabi, UAE. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A Frustratingly Easy Approach for Entity and Relation Extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. In *The Twelfth International Conference on Learning Representations*.

A Experiments

A.1 Baseline Models

We compare the s-LLMs against two categories of foundation LLMs:

1. **Proprietary LLMs:** We use GPT4 (GPT-4o) (Achiam et al., 2023) and Claude (Claude 3-5 Sonnet) (Anthropic, 2024).

2. **Open-source base LLMs:** We include the open-source (vanilla) counterparts of s-LLMs in our evaluation, including Baichuan2 (Baichuan2-7B-Chat) (Yang et al., 2023), and Llama (Llama2-7B-Chat, Llama2-13B-Chat, Llama2-70B-Chat, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct) (Touvron et al., 2023).

In addition, we also compare the performance of LLMs against a fine-tuned **PLM**, i.e., BERT (Devlin et al., 2019) (BERT-base), which consists of an encoder and a span-based classifier on top of the encoder (Zhong and Chen, 2021).

A.2 Evaluation Metrics

We follow prior studies (Lu et al., 2022; Lin et al., 2020) and use strict entity-level micro-F_1 as our evaluation metric, where both the entity boundary and entity type must be correctly predicted.

A.3 Training Environment

We use NVIDIA H100 GPUs for inference and fine-tuning of open-source LLMs and s-LLMs. Our experiments are conducted on a node with two NVIDIA H100 GPUs.

A.4 Datasets

A.4.1 Scientific Domain Datasets

We use four sentence-level datasets, each with a slightly different focus for scientific IE:

1. **MeasEval**[11] (Harper et al., 2021) is a dataset collected from scientific documents from 10 different domains (e.g., agriculture, chemistry and materials science), annotated for four entity types: Quantity, Measured Property, Measured Entity, Qualifier.

2. **SciERC**[12] (Luan et al., 2018) is a dataset collected from the Artificial Intelligence (AI) domain, describing general AI, NLP, Speech Recognition (SR), Machine Learning (ML), and Computer Vision (CV). The entity types: Generic, Material, Method, Metric, OtherScientificTerm and Task.

3. **STEM-ECR**[13] (D'Souza et al., 2020) is a dataset containing scientific abstracts annotated at the sentence-level, covering ten domains (e.g., agriculture and astronomy). Entity types are Material, Data, Process and Method[14].

4. **WLPC**[15] (Kulkarni et al., 2018) is a dataset of technical writing (as opposed to peer-reviewed scientific publications) collected from wet lab protocols for biology and chemistry experiments, providing entity, relation, and event annotations.

The descriptive statistics of all four datasets are listed in Table 4.

[11] https://github.com/harperco/MeasEval

[12] http://nlp.cs.washington.edu/sciIE

[13] https://data.uni-hannover.de/dataset/stem-ecr-v1-0

[14] Although originally there are 7 entity types, we follow previous work (D'Souza et al., 2020) and leave Task, Object, and Results entity types out.

[15] https://github.com/chaitanya2334/WLP-Dataset

Data Split	MeasEval	SciERC	STEM-ECR	WLPC
# Train	542	1,861	942	8.581
# Dev	155	275	118	2,589
# Test	294	551	118	2,861
# Sentences	991	2,687	1,178	14,301
#Word Count	34,779	65,334	25,968	181,908
# Unique Entity Types	4	6	4	18

Table 4: Statistical details of datasets. "#" denotes the number of samples in the specific dataset.

Characteristics of Datasets We note that the first three of the datasets focus on text found in scientific publications, though the scope of the entity detection may be different. For example, the MeasEval dataset focuses on the general concept of quantitative measurements in empirical investigations (e.g., Measured Property). SciERC and STEM-ECR include a combination of specific concepts from the science disciplines as well as general concepts from the scientific method (e.g., Material, Method), although the publication set of SciERC is narrower than that of STEM-ECR. Finally, the WLPC dataset focuses on experimental reports with entity types that differ from the other datasets (given the physical experiment focus), including measure-based (e.g., Numerical, Generic-Measure, Size, Ph, Measure-Type) and science discipline-specific object entities (e.g., Action, Amount, Location).

Of these datasets, only the SciERC was used in the instruction fine-tuning steps for the three models, as the NER task for the UniNER model and the RE task for the IEPile and YAYI-UIE models (Table 6). That is, the data points for the entities and entity types of MeasEval, STEM-ECR, and WLPC datasets were **not seen** during the initial instruction fine-tuning of the s-LLMs.

A.4.2 General Domain Datasets

The descriptive statistics of general domain datasets are given in Table 5.

Dataset	Domain	Type	# Test
CrossNER Politics	Political	9	650
CrossNER Literature	Literary	12	416
CrossNE Music	Musical	13	465
CrossNER AI	AI	14	431
CrossNER Science	Scientific	17	543
CoNLL2003	News	4	3,453

Table 5: The statistical details of the CrossNER dataset. "#" denotes the number of samples in the specific dataset.

Model	Base LLM	Dataset	# Entity Type
IEPile	Llama2-13B & Baichuan2-13B	FabNER (NER) (Kumar and Starly, 2022)	12
		SciERC (RE) (Luan et al., 2018)	4
		SemEval (RE) (Hendrickx et al., 2010)	-
UniNER-7B	Llama2-7B	WLP (Kulkarni et al., 2018)	16
		SoMeSci (Schindler et al., 2021)	14
		SciREX (Jain et al., 2020)	4
		SciERC (Luan et al., 2018)	4
		SOFC (Friedrich et al., 2020)	3
		FabNER (Kumar and Starly, 2022)	12
		DEAL (Grezes et al., 2022)	30
YAYI-UIE	Baichuan2-13B	FabNER (NER) (Kumar and Starly, 2022)	12
		SciERC (RE)	4

Table 6: Statistical details of scientific datasets used in instruction-tuning of IE s-LLMs. "#" denotes the number of entity types in the entity type set. Details are from Zhou et al. (2024).

A.4.3 Benchmark Datasets

Statistical details of scientific datasets used in instruction-tuning of IE (s-LLMs) are given in Table 6. You can find the complete list of datasets in the respective original papers.

A.5 Models and Fine-tuning

For further supervised fine-tuning (SFT) experiments, we use IE s-LLMs (UniNER, IEPile, YAYI-UIE), which are open-source LLMs instruction-tuned for IE tasks and open-source LLMs. Specifically, we employ LoRA (Hu et al., 2022) for parameter-efficient fine-tuning. We follow the previous works for the hyperparameters of SFT (Gui et al., 2024b; Zhou et al., 2024). We set the LoRA rank and alpha parameters to 16 and 32, respectively. The dropout ratio is set to 0.05. The learning rate is set to 5e-5. We limit the input source length to 400 and the target length to 512. The training epoch size is 10, and the batch size is 2.

Baseline BERT-base PLM is fine-tuned utilising the Hugging Face[16] (Wolf et al., 2020) library. The hyperparameters used in the fine-tuning PLM are the batch size of 32, the max length of 128, the learning rate of 1e-5, and 15 epochs of training.

A.6 Zero-Shot and Few-Shot Settings

We conduct zero-shot and few-shot experiments on open-source and proprietary LLMs using the NER prompt of EasyInstruct[17] (Ou et al., 2023). We use random sampling for a few-shot setting, where we select 1 sample from the train set. We set the temperature to 0.0 for results with less variability and set the top probability to 0.95. We use the original prompt templates used in the training of the respective IE s-LLMs in the experiments with these models to align with the setup of the respective NER-specific training regimes.

Prompts We follow EasyInstruct (Ou et al., 2023) in our experiments for open-source and proprietary LLMs. For each dataset, we use its defined entity types and samples (text) from the test set.

IEPile:

```
User: You are an expert in named entity
     recognition. Please extract
    entities that match the schema
    definition from the input. Return
    an empty list if the entity type
    does not exist. Please respond in
    the format of a JSON string.,
    schema: {entity_types}, input: {
    Text}
```

UniNER:

```
User: Text: {Text}
Assistant: I've read this text.
User: What describes {entity_type} in
    the text?
```

YAYI-UIE:

```
User: Text: {Text}
From the given text, extract all the
    entities and types. Please format
    the answer in JSON {{{', '.join(
    entity_types)}: [entities]}}
```

General:

[16] https://huggingface.co
[17] https://github.com/zjunlp/EasyInstruct

```
User: You are a highly intelligent and
    accurate {domain} domain Named-
    entity recognition(NER) system. You
     take Passage as input and your
    task is to recognize and extract
    specific types of {domain} domain
    named entities in that given
    passage and classify into a set of
    following predefined entity types:
{entity_types}
our output format is only [{'E': type
    of entity from predefined entity
    types, 'W': entity in the input
    text},...] form, no other form.
Input: {Text}
```