Open-Source Large Language Models Excel in Named Entity Recognition

Dengya Zhu[1], Sirui Li[2], Nik Thompson[1], and Kok Wai Wong[2]

[1] Discipline of Business Information Systems, School of Management and Marketing, Curtin University, Kent St, Bentley, Western Australia
[2] School of Information Technology, Murdoch University, South St, Murdoch, Western Australia

Abstract. Current state-of-the-art Named Entity Recognition (NER) typically involves fine-tuning transformer-based models like BERT or RoBERTa with annotated datasets, posing challenges in annotation cost, model robustness, and data privacy. An emerging approach uses pre-trained Large Language Models (LLMs) such as ChatGPT to extract entities directly with a few or zero examples, achieving performance comparable to fine-tuned models. However, reliance on the close-source commercial LLMs raises cost and privacy concerns. In this work, we investigate open-source LLMs like Llama2 for NER on local consumer-grade GPUs, aiming to significantly reduce costs compared to cloud solutions while ensuring data security. Experimental results demonstrate competitive NER performance, achieving F1 85.37% on the CoNLL03 dataset and can also be generalised to specific domains, such as scientific texts.

Keywords: Natural language processing · Named entity recognition · Large language models · Open source software · Close source software · Evaluation

1 Introduction

Named Entity Recognition (NER) plays a pivotal role in Natural Language Processing (NLP) by identifying and classifying entities such as names of persons, organisations, locations, dates, and more within unstructured text [8]. Effective NER is crucial for numerous applications in different domains. For example, over 30 million publications in PubMed [23] use NER to extract biomedical entities such as genes, proteins, drugs, and diseases, facilitating downstream tasks.

State-of-the-art (SoTA) NER models [35, 9] are typically supervised fine-tuning BERT-based models. However, fine-tuned NER model have a list of weakness: 1) they always require human-labelled high-quality data with sufficient samples; 2) the labelling process itself is very expensive and labour-intensive; 3) NER labelling is subjective and error-prone—different people may label the same entity differently. For instance, given “Apple iPhone 12", one person may classify the entire phrase as a Product, while another may label “Apple" as an Organisation and “iPhone 12" as a product; 4) a fine-tuned NER model that

Zhu, D., Li, S., Thompson, N., & Wong, K. (2024). Open-Source Large Language Models Excel in Named Entity Recognition. 31st International Conference on Neural Information Processing (ICONIP), Auckland, New Zealand.

performs well on one dataset may not perform as expected due to out-of-domain or out-of-distribution issues; 5) adding a new NER type to an existing dataset is challenging because the entire dataset needs review; 6) acquiring annotated datasets for low-resource languages presents additional challenges [4].

Recently, the launch of Large Language Models (LLMs) marks a transformative era in NLP, demonstrating success in tasks from natural language understanding to question answering across multiple languages [3]. LLM-based NER can effectively alleviate the issues discussed above by following instructions constructed from one or a few sample sentences (known as prompts). While LLMs have been found very successful in many areas of NLP applications, experimental results [10] demonstrated that the NER performance of LLMs such as ChatGPT is still far behind the specifically fine-tuned small language model, with F1 score of ChatGPT's 30% vs 91% (SoTA) model [26, 16]. However, all existing works only evaluate closed-source LLMs [10, 26, 37, 41]. To the best of our knowledge, there is no literature evaluating open-source LLMs for their NER capabilities. This inspires our research question: *"Can open-source LLMs effectively perform NER?"*

Addressing this research question involves three primary concerns: cost, data privacy and data annotation. Compared to commercial and closed-source LLMs like OpenAI's GPTx[3] which require users to upload data via paid APIs, potentially compromising sensitive information [40], open-source LLMs can be freely downloaded and run locally. There are two popular ways to use LLMs: fine-tuning and in-context learning. Fine-tuning also requires many annotated data; however, in-context learning only needs a few examples, known as few-shot learning.

This work explores the ability of open-source LLMs on NER with one-shot and few-shot learning. Specifically, we select three LLMs in our experiments with the consideration of three factors: popularity, model size, and ranked position on the Huggingface open-source leaderboard[4]. We consider LLMs that have more than 7 billion parameters, and released after November of 2022, the release of OpenAI's ChatGPT. More details about model selection can be found in Section 4.1. Our code is available on GitHub[5]. Our contributions are:

- Existing works use close-source LLMs for NER which raise cost and data privacy concerns in specific domains. To the best of our knowledge, our research is the first in the literature to evaluate open-source LLMs for their NER capabilities.
- We compare the performance of different LLMs from one-shot to four-shot scenarios. This approach allows us to evaluate the adaptability and learning efficiency of each LLM, providing insights into their strengths and limitations in handling NER tasks with varying levels of data availability. It opens the door for business domain users to conduct NER projects themselves, without heavily depending on expensive NER experts.

[3] https://platform.openai.com/docs/overview
[4] https://huggingface.co/open-llm-leaderboard
[5] https://github.com/Simon-ozgit/LLM-Eva

Zhu, D., Li, S., Thompson, N., & Wong, K. (2024). Open-Source Large Language Models Excel in Named Entity Recognition. 31st International Conference on Neural Information Processing (ICONIP), Auckland, New Zealand.

– Our experimental datasets include both common domains and specific domain, ensuring that our results are more generalisable to a broader audience.

2 Related Works

2.1 From Fine-tuning to Few-shot/zero-shot Learning

Fine-tuning: Fine-tuning involves adding a layer for different NLP tasks and then fine-tuning the parameters of a pre-trained model to minimise task-specific parameters [28, 7]. One of the extreme approaches for different NLP tasks is to re-fit a task-specific architecture like ELMo, as proposed by [24]. ELMo encodes not only syntactic features from training data but also the semantics of the data in a contextual manner, thereby enhancing overall NLP task performance. However, this approach introduces task-specific parameters and does not effectively leverage pre-trained models. GPT [27] addresses this issue by converting all structured input into token sequences to match its autoregressive decoder-only architecture and then adding an output layer with softmax normalisation. GPT-2 [28] utilises the same fine-tuning strategy, achieved SoTA performance for 7 out of 8 test datasets. BERT [7] also utilises a pre-trained model to initialise each of the downstream NLP task architectures, except for the output layer. It pre-trains a model with unlabelled data for various pre-training tasks; during fine-tuning, all parameters are fine-tuned with annotated data for different NLP tasks [7]. This ensures all NLP tasks share nearly identical architecture, and leverage the weights of the pre-trained model. BERT push the overall GLUE [33] score up absolute 7.7%. To obtain better performance, each task may require thousands or tens of thousands labelled training instances, which is labour intensive and expensive [3].

Few/zero-shot learning, LLM In-context learning, prompting: In-context learning was introduced by OpenAI in 2020 when it released GPT-3, an LLM with 175 billion parameters, trained with 300 billion tokens [3]. As the architecture of LLMs becoming more complex and larger, and trained with more data, fine-tuning such LLMs are becoming harder for both data size and computational power. As estimated by Lambdalabs[6], training GPT-3 may cost $4.6 million US dollars and take 355 years to finish. Importantly, human beings can perform inference with just zero, or a few examples, without any further training/fine-tuning. OpenAI defines the few-shot and zero-shot capabilities of LLMs like GPT-3 as in-context learning, which means LLMs learn broad meta-knowledges and capabilities during training, and automatically apply appropriately learned patterns for different inference tasks, without any task-specific fine-tuning, just with a few examples, or without any examples but simply an instruction (aka prompt). GPT-3 reveals that as model size is increasing, in-context learning capabilities are increasing as wells [3]. Since GPT-3, many LLMs with in-context learning

[6] https://en.wikipedia.org/wiki/GPT-3

Zhu, D., Li, S., Thompson, N., & Wong, K. (2024). Open-Source Large Language Models Excel in Named Entity Recognition. 31st International Conference on Neural Information Processing (ICONIP), Auckland, New Zealand.

capabilities have been developed and released, such as GPT-4 [21], PaLM2 [2], Gemini [1]; and open source models such as Llama [31], Llama2 [32], FlanT5 [5], and other. We omit multi-modal LLMs as they are out of scope of this research.

2.2 Large Language Models for NER

NER capabilities of ChatGPT have been studied via zero-shot settings [37, 41]. Xie et al [41] decomposes NER tasks for specific types with examples, adds syntactic prompting (e.g., part-of-speech tags), employs tool augmentation (using NLP tools or syntactic prompts to provide additional information), and combines these strategies. The augmented prompts are then used to collect responses from ChatGPT, with majority voting determining the final NER results. This method significantly improves over the vanilla zero-shot approach. In the study, Llama-2 consistently performs worse compared to GPT-3.5 and even GPT-3.

GPT-NER [34] is another study aiming at using ChatGPT to extract named entities. This study found that the performance gap between ChatGPT and fine-tuned NER models is due to fine-tuning being a sequence labelling task, while LLM-based NER is an auto-regressive generative model. GPT-NER aims to bridge this gap by converting the sequence labelling task into an auto-regressive task using special tokens to label named entities. GPT-NER also addresses the hallucination [11] issues in LLMs by utilising a self-verification strategy, which asks the LLM itself to verify the extracted entities.

Most research on Information Extraction with LLMs evaluates closed-source models, primarily the ChatGPT series [10, 26, 17, 22, 39].

3 Research Gap

Based on the critical review of the related literature, we address the research gap of NER capabilities of open-source LLMs with the following research questions:

- RQ1: How effective are open-source LLMs at extracting NERs using only one-shot learning?
- RQ2: Do the open-source larger models outperform the smaller models for NER?
- RQ3: How well do open-source LLMs generalise for NER in different domains, such as NER in scientific texts?
- RQ4: How does the number of shots affect the LLMs' NER performance?

This study aims to answer these questions by using popular open-source NER datasets for ease of access and comparison. More details about datasets can be found in Section 5.1. The answers to these questions could have significant impacts on the field of NER for businesses, as the cost of NER could be dramatically reduced by using these one-shot prompts and open-source LLMs. This approach not only eliminates the costly, human-labour-intensive task of training data labelling but also avoids the need for domain-specific training data labelling and fine-tuning, which are both financially costly and time-consuming. In addition, open-source LLMs provide an opportunity to process sensitive information without relying on third-party services that might compromise data privacy.

Zhu, D., Li, S., Thompson, N., & Wong, K. (2024). Open-Source Large Language Models Excel in Named Entity Recognition.
31st International Conference on Neural Information Processing (ICONIP), Auckland, New Zealand.

4 Methodology

We employ one-shot in-context demonstration strategy [3] to extract named entities and evaluate the performance of this approach. Additionally, we extend our prompts to include more examples (from two to four) to address error-prone NER types in this research. For humans, only a brief instruction or a few demonstrations are sufficient to competently perform a new task [3]. This few-shot in-context learning capability of LLMs is desirable because, firstly, collecting thousands or hundreds of thousands of labelled data examples to fine-tune an LLM for every new NLP task is very expensive and limits the efficient usage of LLMs. Secondly, LLMs are pre-trained for broader applications, but fine-tuning tends to narrow down the LLMs' capability and can cause issues such as catastrophic forgetting [29, 15]. Therefore, understanding the one-shot/few-shot NER capability of LLMs is pivotal for effective and efficient NLP applications using LLMs. The instruction utilised in our experiments is, as shown in Figure 1: "Help me extract named entities such as Person, Organisation, Location, and Miscellaneous.".

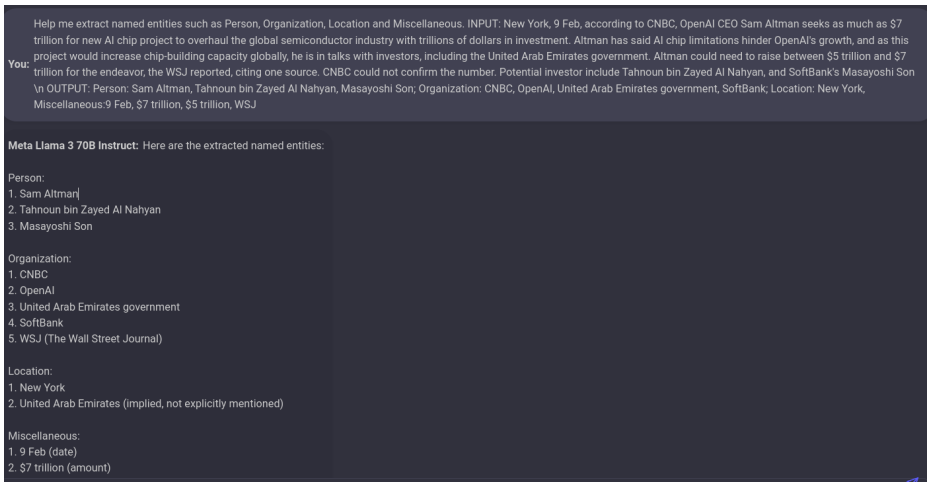


Fig. 1: NER output of Llama-3 70B with one-shot learning.

4.1 LLM Model Selection

We follow considerations of cost, hardware limitations and carbon emissions suggested by [20] to select LLMs for our experiments, constraining the size of the LLMs to the range of approximately 7 billion to 34 billion parameters. Testing models with over 70 billion parameters would require more computational power, be more expensive, necessitate higher hardware configurations, and result in greater carbon emissions. Conversely, although smaller models with fewer than 7 billion parameters have recently shown promising performance on some tasks [18], due to the scaling law and emergent abilities of LLMs [38, 13], we

Zhu, D., Li, S., Thompson, N., & Wong, K. (2024). Open-Source Large Language Models Excel in Named Entity Recognition. 31st International Conference on Neural Information Processing (ICONIP), Auckland, New Zealand.

leave the evaluation of smaller LLMs for future work. Table 1 shows the models used in this paper.

Table 1: Selected Models for NER Experiments

Model	# Parameters	Released Year
Llama2-7B [32]	7 billion[7]	2023
Llama2-13B [32]	13 billion[8]	2023
Llama3[9]	8 billion[10]	2024
Mistral [12]	7 billion[11]	2023
SOLAR [14]	10.7 billion[12]	2023

5 Experiments

5.1 Datasets

We focus on English text only. To facilitate comparisons with experimental results from other research, we used the following three dataset: CoNLL03 [30], OntoNotes5 [25], and SciERC [19]. Details of the datasets are listed in Table 2.

Table 2: Datasets in Experiments

Name	Size (Train/Dev/Test)	NER Type	# of NER Types
CoNLL03	14,041/3,250/3,453	LOC, ORG, PER, MIS	4
OntoNotes5	49,706/13,900/10,348	CARDINAL, DATE, EVENT, FAC, LANGUAGE, LAW, MONEY, PERSON, ORG, LOC, GPE, NORP, ORDINAL, PERCENT, PRODUCT, QUANTITY, TIME, WORK_OF_ART	18
SciERC	1,861/275/551	Task, Method, Evaluation, Metric, Material, Scientific Term, General, Other	8

As point out by [36], annotated datasets contain noise that negatively influences NER performance evaluations. A mis-annotation rate of 5.38% is a significant concern when SoTA performance is as high as approximately 93% in terms of F1 scores [36]. We therefore evaluated our NER performance based on the corrected **test dataset**. We exclude the WNUT17 dataset [6] due to its lack of capitalisation, more heterogeneous NER types, and the fact that its annotation quality has not been thoroughly evaluated [42].

5.2 Metrics

Three metrics are used: precision, recall, and F1 as defined below. Since LLMs might not predict the exact same entity as the answer, we adopt a soft match criterion. For example, given the sentence “Apple iPhone 12 is a new product," the LLM might predict “Apple iPhone 12" as the entity, while the true label is “iPhone 12." Although the prediction is not identical, it is still correct. The soft match criterion accounts for such variations. We define:

Zhu, D., Li, S., Thompson, N., & Wong, K. (2024). Open-Source Large Language Models Excel in Named Entity Recognition.
31st International Conference on Neural Information Processing (ICONIP), Auckland, New Zealand.

– TP: True Positive, the number of NERs correctly assigned to their types
– FP: False Positive, the number of NERs incorrectly assigned to other types
– FN: False Negative, the number of NERs incorrectly rejected by a NER type
– TN: True Negative, the number of NERs corrected rejected by a NER type

$$Precision(p) = \frac{TP}{TP + FP} \quad if\, TP + FP > 0; \quad otherwise\, undefined \tag{1}$$

$$recall(r) = \frac{TP}{TP + FN} \quad if\, TP + FN > 0; \quad otherwise\, undefined \tag{2}$$

$$F1 = \frac{2pr}{p + r} \tag{3}$$

5.3 1-shot Results on CoNLL03

Table 3: Performance of 1-shot on CoNLL03 test dataset. Abbreviations used: 7B (Llama2-7B), 13B (Llama2-13B), L3 (Llama3-8B).

NER Type	Precision					Recall					F1				
	7B	13B	L3	Mistral	SOLAR	7B	13B	L3	Mistral	SOLAR	7B	13B	L3	Mistral	SOLAR
LOC	67.80	78.53	78.11	87.99	**92.41**	77.92	62.54	**95.45**	65.54	66.39	72.51	69.63	**85.91**	75.12	77.27
ORG	67.85	67.84	63.66	57.93	**69.35**	58.11	49.55	**74.54**	42.71	49.40	62.61	57.27	**68.67**	49.17	57.70
PER	96.44	95.24	98.54	98.49	**98.56**	86.51	63.94	**97.49**	85.75	81.69	91.20	76.51	**98.01**	91.68	89.34
Overall	78.17	80.83	80.94	84.17	**88.32**	74.96	59.04	**90.31**	66.39	66.75	76.53	68.24	**85.37**	74.23	76.03

We did not evaluate the performance of MISC due to the inherent differences in how LLMs define MISC. When evaluating the breakdown by entity types (LOC, ORG, PER), a similar trend emerges. For instance, SOLAR achieves the highest precision for identifying all NER types, indicating its ability to minimise false positives. However, Llama3 surpasses it in recall and F1 on all entity types, suggesting a trade-off between precision and recall strategies. According to the leaderboard[13], the highest F1 of CoNLL03 dataset achieved 94.6%, while Llama3 scores 85.37%, indicating a gap of 9.23%. Despite this gap, LLMs demonstrate competitive performance using only one-shot examples.

While analysing the results, we found this gap might also be due to the varying formats generated by LLMs. For example, Llama2-7B sometimes produces predictions formatted as "Person: " and other times as "1. Person:...". Additionally, some predicted entities are on the same line, while others start on a new line with a special symbol. This inconsistency makes it challenging to create a universal regular expression rule that covers all cases, potentially causing our rules to miss some correct answers. Even though we set the temperature to 0.0001, inconsistent formats still occur. Furthermore, LLMs might predict a more accurate NER type that is not among the expected NER types in the dataset. For instance, given the text "CAN / U.S. DOLLAR EXCHANGE RATE : 1.3570.", Llama-7B predicts "Currency: CAN, USD; Exchange Rate: 1.3570". However, "currency" and "exchange rate" are not included in the NER types in the dataset.

[13] https://paperswithcode.com/sota/named-entity-recognition-ner-on-conll-2003

Zhu, D., Li, S., Thompson, N., & Wong, K. (2024). Open-Source Large Language Models Excel in Named Entity Recognition. 31st International Conference on Neural Information Processing (ICONIP), Auckland, New Zealand.

Comparing larger models, such as Llama2-13B and SOLAR, with smaller ones, such as Llama3-8B, underscores the impact of model size on NER performance. Larger models do not consistently achieve higher metrics across all NER types, despite benefiting from increased capacity and potentially richer training data utilisation. In terms of recall and F1 score, Llama2-7B demonstrates better performance (overall 76.53%) than both SOLAR (overall 76.03%) and Llama2-13B (overall 68.24%). The failure of larger models to outperform smaller models could have various factors. One possible explanation could be the quality and quantity of the training data. While a larger model theoretically has the capacity to learn more complex patterns, it also requires a vast amount of high-quality training data to effectively leverage its increased capacity. If the training data for Llama-13B is insufficient or less diverse compared to that of 7B models, it may not be able to generalise as well to unseen data.

5.4 1-shot Results on OntoNotes5

Table 4: Performance of 1-shot on OntoNotes5 test dataset. Abbreviations used: 7B (Llama2-7B), 13B (Llama2-13B), L3 (Llama3).

NER Type	Precision					Recall					F1				
	7B	13B	L3	Mistral	SOLAR	7B	13B	L3	Mistral	SOLAR	7B	13B	L3	Mistral	SOLAR
CARDINAL	2.25	**40.92**	37.35	19.57	8.68	2.11	21.85	7.70	**28.33**	10.82	2.18	**28.49**	12.77	23.15	9.63
DATE	**74.72**	56.35	50.15	54.35	65.30	61.32	29.78	11.37	**62.84**	62.79	**67.36**	38.97	18.54	58.29	64.02
EVENT	8.70	**69.23**	68.75	78.95	64.91	21.05	55.10	19.64	**91.84**	90.24	12.31	61.36	30.56	**84.91**	75.51
FAC	26.85	20.48	**35.29**	5.88	7.14	**54.72**	25.76	18.18	16.67	28.57	**36.02**	22.82	24.00	8.70	11.43
GPE	47.37	53.52	**60.53**	48.77	54.28	70.44	46.30	18.01	70.77	**79.84**	56.65	49.65	27.76	57.74	**64.62**
LANGUAGE	15.38	62.50	**100.00**	58.82	84.21	18.18	62.50	18.18	66.67	**84.21**	16.67	62.50	30.77	62.50	**84.21**
LAW	24.14	33.33	**87.50**	53.57	41.38	41.18	17.24	18.42	**57.69**	54.55	30.43	22.73	30.43	**55.56**	47.06
LOC	10.76	9.28	28.85	13.79	**36.05**	45.95	10.00	10.64	37.74	**63.10**	17.44	9.63	15.54	20.20	**45.89**
MONEY	9.50	40.99	94.12	78.45	**94.16**	18.58	30.28	10.29	**93.57**	92.57	12.57	34.83	18.55	85.35	**93.36**
NORP	4.49	7.71	**23.88**	14.31	7.77	7.84	7.84	7.24	**32.49**	17.04	5.71	7.77	11.11	**19.87**	10.67
ORDINAL	8.86	34.62	**38.89**	25.00	37.33	5.88	19.29	8.28	16.41	**19.44**	7.07	24.77	13.66	19.81	**25.57**
ORG	57.08	38.00	53.58	72.38	**76.67**	61.22	26.04	9.72	66.88	**72.66**	59.07	30.90	16.46	69.52	**74.62**
PERCENT	25.58	54.69	**100.00**	66.43	76.64	8.09	11.40	1.49	**79.17**	57.34	12.29	18.87	2.93	**72.24**	65.60
PERSON	81.17	73.36	92.87	79.82	**95.25**	77.56	42.84	23.68	77.73	**83.01**	79.32	54.09	37.73	78.76	**88.71**
PRODUCT	20.37	42.86	29.41	**43.10**	37.29	35.48	27.78	8.06	**60.98**	59.46	25.88	33.71	12.66	**50.51**	45.83
QUANTITY	65.22	68.97	23.81	51.76	**79.63**	34.09	21.05	13.89	**69.84**	46.24	44.78	32.26	17.54	**59.46**	58.50
TIME	33.85	35.65	**80.00**	37.06	63.64	35.48	30.15	11.76	44.17	**46.39**	34.65	32.67	20.51	40.30	**53.66**
WORK_OF_ART	28.57	17.46	46.67	22.92	**57.03**	47.37	9.65	4.43	60.00	**65.77**	35.64	12.43	8.09	33.17	**61.09**
Overall	50.43	48.00	58.70	54.18	**62.37**	55.35	31.89	13.79	65.50	**67.52**	52.78	38.32	22.34	59.30	**64.84**

Table 4 shows that each model exhibits strengths and weaknesses depending on the NER type evaluated. For instance, Mistral has the best F1 score for EVENT (84.91%), whereas SOLAR achieves the best F1 score for MONEY (93.36%). Overall, SOLAR demonstrates the highest precision, recall, and F1 scores. In terms of precision, Llama3 achieves the highest precision for the most NER types (8 out of 18), but its recall is relatively low. SOLAR and Mistral achieves the highest recall across the most NER types (9 out of 18). For F1 scores, SOLAR leads with the highest F1 scores in 10 out of 18 NER types.

The performance on the OntoNotes5 dataset is significantly lower than on the CoNLL03 dataset, likely due to the increased difficulty presented by the greater number of NER types in OntoNotes5. According to the leaderboard[14], the

[14] https://paperswithcode.com/sota/named-entity-recognition-ner-on-ontonotes-v5

Zhu, D., Li, S., Thompson, N., & Wong, K. (2024). Open-Source Large Language Models Excel in Named Entity Recognition. 31st International Conference on Neural Information Processing (ICONIP), Auckland, New Zealand.

highest F1 score on the OntoNotes5 dataset achieved without extra training data is 91.9%, while SOLAR scores 64.84%, indicating a substantial gap of 27.06%. This limitation might be due to a single example being insufficient to instruct the model to distinguish complex NER types. To investigate this, we increase the number of shots in Section 5.6. Again, larger models, such as Llama2-13B, do not consistently outperform smaller models.

5.5 1-shot Results on SciERC

Table 5: Performance of 1-shot on sciERC test dataset. Abbreviations used: 7B (Llama2-7B), 13B (Llama2-13B), L3 (Llama3).

NER Type	Precision					Recall					F1				
	7B	13B	L3	Mistral	SOLAR	7B	13B	L3	Mistral	SOLAR	7B	13B	L3	Mistral	SOLAR
Method	76.52	70.29	66.56	**83.65**	63.80	27.98	52.34	68.15	**78.24**	61.17	40.97	60.00	67.34	**80.85**	62.46
Task	**75.53**	71.24	56.12	54.74	70.83	31.70	53.69	68.32	64.60	**71.20**	44.65	61.24	61.62	59.26	**71.02**
Other Scientific Term	4.00	4.74	**39.27**	34.88	24.35	1.06	5.19	**54.09**	43.30	31.38	1.68	4.96	**45.50**	38.63	27.42
Material	22.22	56.82	**87.60**	24.14	64.91	4.76	45.87	**80.30**	25.93	69.16	7.84	50.76	**83.79**	25.00	66.97
Generic	0.00	1.11	**4.46**	3.92	1.92	0.00	0.94	**5.68**	4.12	2.15	0.00	1.02	**5.00**	4.02	2.03
Metric	30.00	67.39	**77.05**	8.57	76.79	4.92	58.49	**87.04**	8.33	78.18	8.45	62.63	**81.74**	8.45	77.48
Overall	44.58	42.69	**53.02**	48.39	48.24	14.33	36.27	**61.59**	52.36	52.05	21.69	39.22	**56.98**	50.30	50.07

From Table 5, it is obvious that Llama3 consistently performs well in terms of precision, recall, and F1 across most categories. SOLAR and Mistral also show strong performance in specific categories, with SOLAR excelling in precision and Mistral in certain F1 scores. According to the leaderboard[15], the highest F1 score on the sciERC dataset is 72.4% . In comparison, Llama3 scores 56.98%, indicating a substantial gap of 15.42%. While analysing the results, we also found this gap might be due to the varying formats generated by LLMs, which has been analysed in Section 5.3. This performance is achieved with only one-shot examples, which demonstrates the effectiveness of LLMs in specific domains with minimal examples.

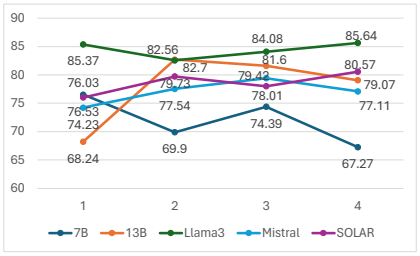
5.6 Few-shot Results

In Section 5.4 and Section 5.5, we hypothesize that the gap between SoTA and our results might be due to insufficient examples. To verify the hypothesis, we add more examples to the prompt. All examples are randomly selected from the training or dev datasets and contain as many NER types as possible.

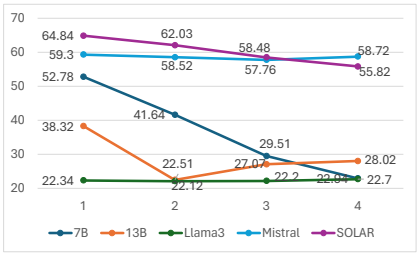
The results for the CoNLL03 dataset (Figure 2a) reveal that the F1 of the models show distinct trends. Except for the Llama-7B model, all models perform better with four shots compared to one shot, though there is a slight drop at two or three shots for some models. For example, Mistral's performance increases from one shot (76.53%) to four shots (77.11%), but its highest score is at three shots. Similarly, Llama3 and SOLAR achieve their highest scores at four shots, while Llama2-13B peaks at two shots. Interestingly, Llama-7B model performs best at one shot but records its lowest score at four shots. This trend might

[15] https://paperswithcode.com/sota/named-entity-recognition-ner-on-scierc

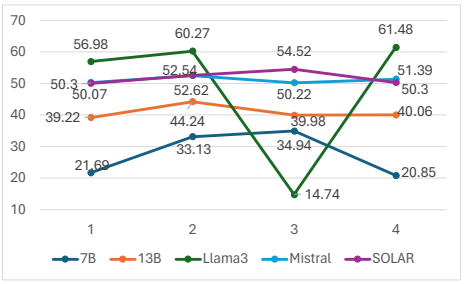
Zhu, D., Li, S., Thompson, N., & Wong, K. (2024). Open-Source Large Language Models Excel in Named Entity Recognition. 31st International Conference on Neural Information Processing (ICONIP), Auckland, New Zealand.



(a) CoNLL03



(b) OntoNotes5



(c) sciERC

Fig. 2: Overall F1 results across different datasets. The x-axis represents the number of shots, ranging from 1 to 4, and the y-axis represents the F1 value.

be due to the varying complexity of the NER tasks, where additional examples introduce variability that can momentarily hinder performance. This result is also reflected in the OntoNotes5 dataset, a more complex dataset with more NER types, where results with one shot outperform those with more shots. SOLAR and Llama-7B have the lowest scores at four shots.

From Figure 2c, we can see that increasing from one shot to two shots slightly improves performance on domain-specific datasets. Some models continue to improve with three shots, but performance with four shots might even be lower than with one shot.

In summary, for common NER types, increasing the number of shots might enhance performance to some extent. However, if the NER types are complex and not common, additional examples also introduce variability that can momentarily hinder performance.

6 Conclusion

This paper presents a comprehensive evaluation of the NER capabilities of open-source LLMs, addressing a gap in the existing literature that has predominantly focused on closed-source models. Our experiments reveal that open-source LLMs perform competitively, particularly in extracting common entity types such as

Zhu, D., Li, S., Thompson, N., & Wong, K. (2024). Open-Source Large Language Models Excel in Named Entity Recognition. 31st International Conference on Neural Information Processing (ICONIP), Auckland, New Zealand.

Person, Organisation, and Location across general domains. Notably, larger models do not always guarantee superior performance, especially when dealing with more complex or less frequent NER types. While an increase in training data can slightly improve performance for common entities, inconsistencies in accuracy emerge when handling diverse and less frequent NERs. These findings suggest that model performance is influenced by both entity complexity and data availability.

In the future, addressing format inconsistencies in LLM-generated outputs and experimenting with more diverse prompts will be critical for enhancing NER performance across various domains. We will expand our work by exploring more NLP tasks to assess the broader capabilities of open-source LLMs. This study opens the door for future work focused on optimising LLMs in domain-specific tasks and developing strategies to handle rare and complex entities more effectively, thereby enhancing their utility across various applications.

References

1. Anil, R., Borgeaud, S., et al.: Gemini: A family of highly capable multimodal models. CoRR **abs/2312.11805** (2023).
2. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J.H., Shafey, L.E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., Ruder, S., Tay, Y., Xiao, K., Xu, Y., Zhang, Y., Ábrego, G.H., Ahn, J., Austin, J., Barham, P., Botha, J.A., Bradbury, J., Brahma, S., Brooks, K., Catasta, M., Cheng, Y., Cherry, C., Choquette-Choo, C.A., Chowdhery, A., Crepy, C., Dave, S., Dehghani, M., Dev, S., Devlin, J., Díaz, M., Du, N., Dyer, E., Feinberg, V., Feng, F., Fienber, V., Freitag, M., Garcia, X., Gehrmann, S., Gonzalez, L., et al.: Palm 2 technical report. CoRR **abs/2305.10403** (2023).
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901
4. Buaphet, W., Udomcharoenchaikit, C., Limkonchotiwat, P., Rutherford, A., Nutanong, S.: Thai nested named entity recognition corpus. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 1473–1486
5. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V.Y., Huang, Y., Dai, A.M., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling instruction-finetuned language models. CoRR **abs/2210.11416** (2022).
6. Derczynski, L., Nichols, E., van Erp, M., Limsopatham, N.: Results of the WNUT2017 shared task on novel and emerging entity recognition. In: Derczynski, L., Xu, W., Ritter, A., Baldwin, T. (eds.) Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017. pp. 140–147. Association for Computational Linguistics (2017).
7. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of

Zhu, D., Li, S., Thompson, N., & Wong, K. (2024). Open-Source Large Language Models Excel in Named Entity Recognition. 31st International Conference on Neural Information Processing (ICONIP), Auckland, New Zealand.

the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (2019).
8. Dhiaf, M., Jemni, S.K., Kessentini, Y.: Docner: A deep learning system for named entity recognition in handwritten document images. In: Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part VI 28. pp. 239–246. Springer
9. Ding, N., Xu, G., Chen, Y., Wang, X., Han, X., Xie, P., Zheng, H., Liu, Z.: Few-nerd: A few-shot named entity recognition dataset. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. pp. 3198–3213. Association for Computational Linguistics (2021).
10. Han, R., Peng, T., Yang, C., Wang, B., Liu, L., Wan, X.: Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. CoRR **abs/2305.14450** (2023).
11. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12), 248:1–248:38 (2023).
12. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b. CoRR **abs/2310.06825** (2023).
13. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. CoRR **abs/2001.08361**
14. Kim, D., Park, C., Kim, S., Lee, W., Song, W., Kim, Y., Kim, H., Kim, Y., Lee, H., Kim, J., Ahn, C., Yang, S., Lee, S., Park, H., Gim, G., Cha, M., Lee, H., Kim, S.: SOLAR 10.7b: Scaling large language models with simple yet effective depth up-scaling. CoRR **abs/2312.15166** (2023).
15. Kotha, S., Springer, J.M., Raghunathan, A.: Understanding catastrophic forgetting in language models via implicit inference. CoRR **abs/2309.10105** (2023).
16. Laskar, M.T.R., Bari, M.S., Rahman, M., Bhuiyan, M.A.H., Joty, S., Huang, J.X.: A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. In: Rogers, A., Boyd-Graber, J.L., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023. pp. 431–469. Association for Computational Linguistics (2023).
17. Li, B., Fang, G., Yang, Y., Wang, Q., Ye, W., Zhao, W., Zhang, S.: Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. CoRR **abs/2304.11633** (2023).
18. Li, Y., Bubeck, S., Eldan, R., Giorno, A.D., Gunasekar, S., Lee, Y.T.: Textbooks are all you need II: phi-1.5 technical report. CoRR **abs/2309.05463** (2023).
19. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.: Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. pp. 3219–3232. Association for Computational Linguistics (2018).

Zhu, D., Li, S., Thompson, N., & Wong, K. (2024). Open-Source Large Language Models Excel in Named Entity Recognition.
31st International Conference on Neural Information Processing (ICONIP), Auckland, New Zealand.

20. Manuvinakurike, R., Sahay, S., Manepalli, S., Nachman, L.: Zero-shot conversational summarization evaluations with small large language models. CoRR **abs/2311.18041** (2023).
21. OpenAI: GPT-4 technical report. CoRR **abs/2303.08774** (2023).
22. Pang, C., Cao, Y., Ding, Q., Luo, P.: Guideline learning for in-context information extraction. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023. pp. 15372–15389. Association for Computational Linguistics (2023).
23. Perera, N., Dehmer, M., Emmert-Streib, F.: Named entity recognition and relation detection for biomedical information extraction. Frontiers in cell and developmental biology **8**, 673
24. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Walker, M.A., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics (2018).
25. Pradhan, S., Moschitti, A., Xue, N., Ng, H.T., Björkelund, A., Uryupina, O., Zhang, Y., Zhong, Z.: Towards robust linguistic analysis using ontonotes. In: Hockenmaier, J., Riedel, S. (eds.) Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013. pp. 143–152. ACL
26. Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., Yang, D.: Is chatgpt a general-purpose natural language processing task solver? In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023. pp. 1339–1384. Association for Computational Linguistics (2023).
27. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training
28. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9
29. Ren, W., Li, X., Wang, L., Zhao, T., Qin, W.: Analyzing and reducing catastrophic forgetting in parameter efficient tuning. CoRR **abs/2402.18865** (2024).
30. Sang, E.F.T.K., Meulder, F.D.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Daelemans, W., Osborne, M. (eds.) Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003. pp. 142–147. ACL
31. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. CoRR **abs/2302.13971** (2023).
32. Touvron, H., Martin, L., et al.: Llama 2: Open foundation and fine-tuned chat models. CoRR **abs/2307.09288** (2023).
33. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: Linzen, T., Chrupala, G., Alishahi, A. (eds.) Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018. pp. 353–355. Association for Computational Linguistics (2018).

Zhu, D., Li, S., Thompson, N., & Wong, K. (2024). Open-Source Large Language Models Excel in Named Entity Recognition. 31st International Conference on Neural Information Processing (ICONIP), Auckland, New Zealand.

34. Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., Wang, G.: GPT-NER: named entity recognition via large language models. CoRR **abs/2304.10428** (2023).
35. Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F., Tu, K.: Automated concatenation of embeddings for structured prediction. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. pp. 2643–2660. Association for Computational Linguistics (2021).
36. Wang, Z., Shang, J., Liu, L., Lu, L., Liu, J., Han, J.: Crossweigh: Training named entity tagger from imperfect annotations. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. pp. 5153–5162. Association for Computational Linguistics (2019).
37. Wang, Z., Zhao, Z., Chen, Z., Ren, P., de Rijke, M., Ren, Z.: Generalizing few-shot named entity recognizers to unseen domains with type-related features. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023. pp. 2228–2240. Association for Computational Linguistics (2023).
38. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W.: Emergent abilities of large language models. Trans. Mach. Learn. Res. **2022**
39. Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., Jiang, Y., Han, W.: Zero-shot information extraction via chatting with chatgpt. CoRR **abs/2302.10205** (2023).
40. Wu, X., Duan, R., Ni, J.: Unveiling security, privacy, and ethical concerns of chatgpt. Journal of Information and Intelligence **2**(2), 102–115
41. Xie, T., Li, Q., Zhang, J., Zhang, Y., Liu, Z., Wang, H.: Empirical study of zero-shot NER with chatgpt. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023. pp. 7935–7956. Association for Computational Linguistics (2023).
42. Zhu, Y., Ye, Y., Li, M., Zhang, J., Wu, O.: Investigating annotation noise for named entity recognition. Neural Comput. Appl. **35**(1), 993–1007 (2023).

Zhu, D., Li, S., Thompson, N., & Wong, K. (2024). Open-Source Large Language Models Excel in Named Entity Recognition.
31st International Conference on Neural Information Processing (ICONIP), Auckland, New Zealand.