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.

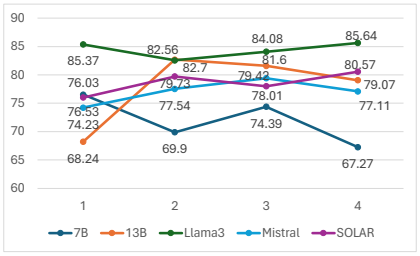
5.6 Few-shot Results

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.

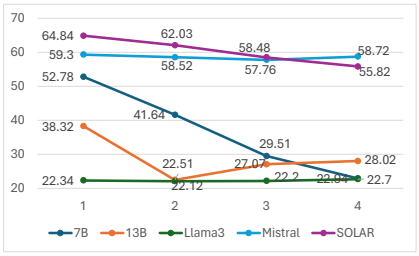
The results for the CoNLL03 dataset (Figure 2a) reveal that the F1 of the models show distinct trends. Except for the Llama-7B model, all models perform better with four shots compared to one shot, though there is a slight drop at two or three shots for some models. For example, Mistral's performance 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

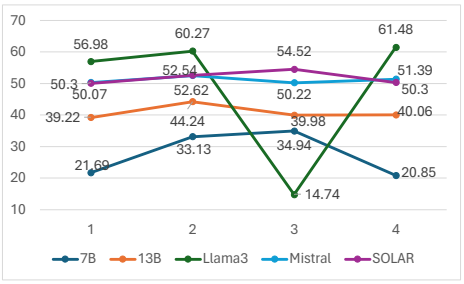
Zhu, D., Li, S., Thompson, N., & Wong, K. (2024). Open-Source Large Language Models Excel in Named Entity Recognition. 31st International Conference on Neural Information Processing (ICONIP), Auckland, New Zealand.



(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.