Improving Named Entity Recognition for Low-Resource Languages Using Large Language Models: A Ukrainian Case Study

Vladyslav Radchenko[1,2], Nazarii Drushchak[1,2]
[1]Ukrainian Catholic University
[2] Softserve Inc
{radchenko, drushchak}.pn@ucu.edu.ua

Abstract

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP), yet achieving high performance for low-resource languages remains challenging due to limited annotated data and linguistic complexity. Ukrainian exemplifies these issues with its rich morphology and scarce NLP resources. Recent advances in Large Language Models (LLMs) demonstrate their ability to generalize across diverse languages and domains, offering promising solutions without extensive annotations. This research explores adapting state-of-the-art LLMs to Ukrainian through prompt engineering, including chain-of-thought (CoT) strategies, and model refinement via Supervised Fine-Tuning (SFT). Our best model achieves **0.89 $\mathbf{F}_1$** on the NER-UK 2.0 benchmark, matching the performance of advanced encoder-only baselines. These findings highlight practical pathways for improving NER in low-resource contexts, promoting more accessible and scalable language technologies.

1 Introduction and Motivation

Accurate identification of named entities underpins a wide range of NLP applications, including information extraction, question answering, and data anonymization, particularly in privacy-sensitive domains such as healthcare, legal document processing, and finance (Keraghel et al., 2024). However, developing robust NER systems for low-resource languages, such as Ukrainian, remains challenging due to the scarcity of annotated datasets and the complexity of linguistic features (Chaplynskyi and Romanyshyn, 2024).

Traditional NER approaches, including rule-based methods and early deep learning models, rely on large annotated corpora, which are difficult to obtain for low-resource languages (Li et al., 2022; Brandsen et al., 2020). Ukrainian's rich morphology and free word order further complicate direct adaptation from resource-rich languages (Chaplynskyi and Romanyshyn, 2024; Artetxe et al., 2020), leaving a significant performance gap.

Recent advances in LLMs offer promising solutions for low-resource NER through zero-shot and few-shot learning, leveraging large-scale pretraining to operate with minimal task-specific data (Shen et al., 2023; Wang et al., 2025). Techniques such as CoT prompting (Wei et al., 2022b) and SFT (Wei et al., 2022a; Keloth et al., 2024) further enhance adaptability to linguistic complexity. In this study, we also evaluate state-of-the-art encoder-only models as competitive baselines to assess whether LLM-based approaches offer measurable gains. Our goal is to address data scarcity in Ukrainian NER and contribute to bridging the performance gap between low- and high-resource languages (Monajatipoor et al., 2024).

The remainder of this paper is structured as follows. Section 2 reviews related literature. Section 3 defines research gaps and study objectives. Section 4 describes the dataset. Section 5 outlines the methodology, including model selection, experimental setup, and evaluation. Section 6 presents and analyzes the results. Section 7 summarizes findings and suggests future directions. Section 8 discusses limitations, and covers ethical considerations.

2 Related Work

2.1 NER Fundamentals

Early NER systems relied on rule-based methods using manually created rules, dictionaries, and regular expressions. Though effective for structured texts, these systems lacked flexibility and scalability across diverse domains and languages (Aliwy et al., 2021). Feature-based machine learning approaches, including Conditional Random Fields (CRFs) and Support Vector Machines (SVMs), reduced manual rule creation by leveraging linguis-

Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), pages 27–35
July 31 - August 1, 2025

tic features but still required extensive annotated datasets (Li et al., 2022).

The adoption of deep learning transformed NER methods. Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks (Sherstinsky, 2020), automated feature extraction and enhanced performance. Transformer-based encoder-only architectures, notably BERT (Devlin et al., 2019), further improved results through self-attention mechanisms (Vaswani et al., 2017), setting new benchmarks. However, these models are highly dependent on high-quality, resource-rich data to effectively generalize across varied linguistic contexts.

2.2 NER in Low-Resource Languages

Low-resource languages like Ukrainian pose challenges due to limited annotated corpora, complex morphology, and flexible syntax. These characteristics demand expert annotation and make the development of robust models particularly difficult (Brandsen et al., 2020). To mitigate the need for extensive labeled data, researchers have explored alternative strategies such as transfer learning, data augmentation, zero-shot prompting, and active learning (Keraghel et al., 2024).

The most comprehensive publicly available resource is NER-UK 2.0 (Chaplynskyi and Romanyshyn, 2024), a manually annotated dataset covering a wide range of genres and entity types. Other initiatives, such as a news-focused dataset described in (Makogon and Samokhin, 2022), have not been released publicly, limiting their utility for reproducible research. Automatically annotated corpora—such as POLYGLOT-NER (Venkatasubramanian and Ye, 2015), WikiANN (Pan et al., 2017), and Ukr-Synth2[1]—offer broader coverage but are constrained by limited entity schemas and lack human verification. The SlavNER corpus (Piskorski et al., 2024) includes high-quality manual annotations for Ukrainian, though it is restricted to five entity types and Wikipedia-derived text. Overall, these resources provide useful foundations, but vary in quality, genre diversity, and annotation scope—highlighting the need for a robust, publicly available dataset with rich entity coverage.

2.3 Large Language Models and NER

LLMs such as GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023) have demonstrated strong performance in NER, particularly in low-resource settings. Pre-trained on large-scale corpora, these models generalize well across domains and require minimal task-specific supervision. Their ability to perform NER in zero-shot and few-shot scenarios makes them especially suitable for languages with limited annotated data (Brown et al., 2020; Ji, 2023; Hu et al., 2024; Monajatipoor et al., 2024; Li and Zhang, 2024; Shen et al., 2023).

In zero-shot settings, LLMs extract entities based on natural language instructions, while few-shot setups incorporate a small number of labeled examples to improve accuracy. Methods like GPT-NER (Wang et al., 2025) and PromptNER (Shen et al., 2023) showcase the effectiveness of prompt-based approaches across both low-resource and domain-specific NER tasks.

SFT and prompt engineering improve LLM performance by aligning model behavior with task-specific prompts, showing strong results in domains like biomedical NER (Keloth et al., 2024). While challenges remain, such as high computational cost and prompt sensitivity, LLMs have proven effective in Ukrainian NLP tasks (Paniv et al., 2024), making them promising for low-resource NER.

3 Research Gaps and Objectives

Despite progress, Ukrainian NER faces key challenges: limited high-quality annotated data, underexplored use of LLMs, and heavy reliance on proprietary models, which restricts transparency. In addition, the absence of standardized benchmarks hinders consistent evaluation and comparison.

To address these gaps, this study pursues the following objectives:

- Investigate the effectiveness of LLMs for Ukrainian NER under prompt-based and supervised fine-tuning scenarios.
- Benchmark open-source LLMs against proprietary models to assess their viability in low-resource settings.
- Propose standardized evaluation pipeline for LLMs.

4 Dataset Overview

Given the limitations of existing resources, we select NER-UK 2.0 (Chaplynskyi and Romanyshyn, 2024) as the primary benchmark for this study. It is the largest public manually annotated Ukrainian

[1] https://huggingface.co/datasets/ukr-models/Ukr-Synth

NER corpus, comprising 560 texts and 21,993 entities across 13 categories. The dataset includes diverse genres—such as news, social media, legal documents, and procurement contracts, and follows the widely adopted Inside-Outside-Beginning labeling scheme.

NER-UK 2.0 offers comprehensive entity coverage but has limitations like domain bias, class imbalance (e.g., frequent `PERS` and `ORG` vs. rare `DOC` and `TIME`), and subjective annotation challenges (e.g. `MISC`). Despite these, it remains invaluable for Ukrainian NER research.

5 Methodology

5.1 Experiments Set Up

A series of experiments will be conducted to evaluate the performance of the LLM models under different conditions, structured as follows:

- **Encoder-only Model Fine-tuning.** Establishes a robust baseline using state-of-the-art encoder models, providing a point of comparison for LLM-based approaches. Training is conducted via `spaCy`[2] pipeline.

- **Zero-shot, Few-shot, and CoT Prompting.** Assesses model performance with minimal annotated data, reflecting realistic low-resource scenarios. Inference is performed using `vLLM`[3] for scalable decoding.

- **LLM Supervised Fine-tuning.** Assesses fine-tuned LLMs against encoder baselines, with a focus on rare entity types. Fine-tuning is carried out using `Unsloth`[4] with LoRA adapters for parameter-efficient training, and inference is performed using `Transformers`[5].

5.2 Model Selection

We selected top-performing LLMs from diverse architectures, including high-ranking open-source models from the Hugging Face Open LLM Leaderboard[6] and proprietary models accessed via APIs. To manage computational constraints, open-source models were limited to 27 billion parameters, ensuring a balanced comparison. A full list of selected models is provided in Appendix A.

[2] https://spacy.io/
[3] https://docs.vllm.ai/
[4] https://unsloth.ai/
[5] https://huggingface.co/docs/transformers
[6] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

To establish meaningful baselines, we trained prominent encoder-only models on the Ukrainian NER dataset. These included GLiNER (Zaratiana et al., 2024), XLM-RoBERTa (Conneau et al., 2019), Modern BERT (Warner et al., 2024) variants, as well as other transformer-based models pre-trained on multilingual or domain-specific corpora relevant to Ukrainian NER. Such models offer strong performance in resource-efficient setups and serve as reliable benchmarks to evaluate the added value of LLM-based approaches.

5.3 Evaluation

This study uses the **F1-score** as the primary evaluation metric. Following the NER-UK 2.0 (Chaplynskyi and Romanyshyn, 2024) paper, employing entity-level evaluation.

To assess model performance under different validation levels, we define three evaluation stages:

- **Bronze.** Raw model output without any validation or cleaning.

- **Silver.** Light cleaning of LLM outputs, removing hallucinations and correcting word variants via char-level similarity[7].

- **Gold.** Rule-based filtering enforcing constraints like disallowing person entities that begin with lowercase letters or are pronouns[8].

The code and experiments are available[9].

6 Results and Discussion

6.1 Encoder-Only Model Fine-Tuning

Encoder-based models show consistent performance, with F_1 scores ranging **from 0.855 to 0.890** (Appendix B). During this study, we identified and corrected a training issue in the previously released `uk-ner-web-trf-13class`, where the test set was inadvertently used used as evaluation set to define best model. The model was retrained with the appropriate validation setup for fair comparison.

`ModernBERT-large` underperforms, reaching 0.762 F_1, likely due to its monolingual architecture and limited exposure to Ukrainian. The best performance is achieved by `roberta-large-NER` with **0.890** F_1, showing strong results across both frequent (`PERS`, `ORG`) and less frequent (`ART`, `JOB`) entity types, indicating robust generalization.

[7] Char n-gram cosine similarity aligns noisy spans with valid input.
[8] Pronouns are detected using POS tags from `stanza`.
[9] https://github.com/pofce/NER-Ukrainian-LLMs

6.2 Zero-Shot, Few-Shot, and CoT Prompting

Few-shot prompting consistently outperforms zero-shot, confirming the effectiveness of minimal in-context learning. CoT prompting does not yield consistent improvements, suggesting its limited value for span-based tasks. Full results are available in Appendix C.

Post-processing significantly improves output quality; moving from Bronze to Gold evaluation often yields substantial F_1 gains, indicating that LLMs frequently generate near-correct predictions that benefit from light normalization.

While larger models generally perform better, architecture and pretraining quality remain critical. Notably, open-source models like Gemma-3-27B-IT reach **0.71** F_1, closing the gap with proprietary models such as GPT-4. However, this performance comes at the cost of added complexity. In contrast, generalist models like `gliner` achieve up to **0.67** F_1 (Appendix D) with minimal setup, highlighting a trade-off between performance and usability. [10]

6.3 LLM Supervised Fine-Tuning

Supervised fine-tuning of LLMs yields performance comparable to encoder-only baselines. For instance, `Gemma-3-27B-IT` reaches **0.888** F_1, closely aligning with `roberta-large-NER` (Appendix F). However, gains are limited on low-resource categories such as `TIME`, `MISC`, and `DOC`, indicating that increased model capacity alone does not resolve data sparsity challenges.

All LLMs were fine-tuned with minimal hyperparameter tuning for consistency and efficiency (Appendix E). While fine-tuned LLMs remain competitive, their marginal improvements relative to computational cost highlight the need for more efficient and targeted approaches for low-resource NER.

7 Conclusion and Future Work

LLMs demonstrate strong performance for Ukrainian NER under minimal supervision, particularly in few-shot settings. However, this comes at the cost of increased computational demands and system complexity. In contrast, generalist models like `gliner`, while less accurate, offer a more efficient and accessible alternative.

Supervised fine-tuning of LLMs yields results comparable to encoder-only baselines but provides limited improvement on low-resource entity types and requires significantly more resources.

`roberta-large-NER` emerged as the best-performing model on the NER-UK 2.0 benchmark, establishing a new state-of-the-art. A full side-by-side comparison of top models from each approach is provided in Appendix G.

Model	Experiment	F1 Score
roberta-large-NER	Fine-tuning	0.890
Gemma-3-27B-IT	Fine-tuning	0.888
GPT-4o	Zero-shot	0.724
Gemma-3-27B-IT	Few-shot	0.712
GLiNER	Zero-shot	0.670

Table 1: Best-Performing Models Across Approaches

Future work will explore adapting LLMs into encoder-style architectures for more efficient token-level prediction and reinforcement learning from human feedback tuning techniques. We also plan to annotate the social media portion of UberText 2.0 (Chaplynskyi, 2023) using the best-performing model to create a silver-standard NER dataset.

Limitations and Ethical Considerations

This study acknowledges several limitations:

- The analysis focused on open-source models under 27B parameters, and proprietary models were minimally considered due to limited access.
- Prominent LLM-based NER techniques were not extensively applied due to time and resource constraints.
- LLMs were treated as generative models; integration into encoder-style architectures for token-level prediction remains unexplored and may offer benefits in span-based tasks.
- All experiments were based on a single dataset.

In this study, no personally identifiable information was used. ChatGPT[11] was used to paraphrase and improve the textual clarity during the writing process.

[10] Prompt templates and code are available at https://github.com/pofce/NER-Ukrainian-LLMs/tree/main/experiments/prompting

[11] https://chatgpt.com/

References

Ahmed Aliwy, Ayad Abbas, and Ahmed Alkhayyat. 2021. Nerws: Towards improving information retrieval of digital library management system using named entity recognition and word sense. *Big Data and Cognitive Computing*, 5:1–16.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 7375–7388. Association for Computational Linguistics.

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577. European Language Resources Association.

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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: A corpus of modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Dmytro Chaplynskyi and Mariana Romanyshyn. 2024. Introducing NER-UK 2.0: A rich corpus of named entities for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 23–29. ELRA and ICCL.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820.

Bin Ji. 2023. Vicunaner: Zero/few-shot named entity recognition using vicuna. *Preprint*, arXiv:2305.03253.

Vipina K Keloth, Yan Hu, Qianqian Xie, Xueqing Peng, Yan Wang, Andrew Zheng, Melih Selek, Kalpana Raja, Chih Hsuan Wei, Qiao Jin, Zhiyong Lu, Qingyu Chen, and Hua Xu. 2024. Advancing entity recognition in biomedicine via instruction tuning of large language models. *Bioinformatics*, 40(4):btae163.

Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. Recent advances in named entity recognition: A comprehensive survey and comparative study. *Preprint*, arXiv:2401.10825.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Mingchen Li and Rui Zhang. 2024.

Iuliia Makogon and Igor Samokhin. 2022. Targeted sentiment analysis for ukrainian and russian news articles.

Masoud Monajatipoor, Jiaxin Yang, Joel Stremmel, Melika Emami, Fazlolah Mohaghegh, Mozhdeh Rouhsedaghat, and Kai-Wei Chang. 2024. Llms in biomedicine: A study on clinical named entity recognition. *Preprint*, arXiv:2404.07376.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Yurii Paniv, Artur Kiulian, Dmytro Chaplynskyi, Mykola Khandoga, Anton Polishko, Tetiana Bas, and Guillermo Gabrielli. 2024. Benchmarking multimodal models for ukrainian language understanding across academic and cultural domains. *Preprint*, arXiv:2411.14647.

Jakub Piskorski, Michał Marcińczuk, and Roman Yangarber. 2024. Cross-lingual named entity corpus for Slavic languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4143–4157, Torino, Italia. ELRA and ICCL.

Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023. PromptNER: Prompt locating and typing for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12492–12507, Toronto, Canada. Association for Computational Linguistics.

Alex Sherstinsky. 2020. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, , et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Suresh Venkatasubramanian and Jieping Ye. 2015. *Proceedings of the 2015 SIAM International Conference on Data Mining (SDM)*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

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. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

A. Model Sizes

Model	**Number of Parameters**	**Model Category**
gpt-4o-2024-11-20	-	Proprietary LLM
Gemma-3-27B-IT	27.4B	Open-Source LLM
Gemma-2-27B-IT	27.2B	Open-Source LLM
Gemma-2-9B-IT	9.2B	Open-Source LLM
Phi-4	14.7B	Open-Source LLM
Qwen-2.5-14B-Instruct	14.8B	Open-Source LLM
Qwen-2.5-7B-Instruct	7.6B	Open-Source LLM
DeepSeek-R1-Distill-Qwen-14B	14.8B	Open-Source LLM
Gemma-2-2B-IT	2.6B	Open-Source LLM
Qwen-2.5-3B-Instruct	3.0B	Open-Source LLM
Llama-3.2-3B-Instruct	3.2B	Open-Source LLM
Phi-3-mini-4k-instruct	3.8B	Open-Source LLM
Llama-3.1-8B-Instruct	8.3B	Open-Source LLM
Aya-expanse-8b	8.0B	Open-Source LLM
Aya-101	13.0B	Open-Source LLM
roberta-large-NER	561M	Encoder-only
xlm-roberta-large	561M	Encoder-only
NuNER-Zero	449M	Encoder-only
Modern-BERT-large	396M	Encoder-only
gliner-multi-v2.1	209M	Encoder-only
gliner-multi-pii-v1	209M	Encoder-only
uk-ner-web-trf-13class	110M	Encoder-only

B. Final Results on Encoder-Only Model Tuning

Entity	**roberta-large-NER**	**xlm-roberta-large**	**gliner-multi-v2.1**	**Modern-BERT-large**	**uk-ner-web-trf-13class**
JOB	**0.699**	0.689	**0.699**	0.470	0.696
PERIOD	0.743	0.742	0.712	0.596	**0.769**
QUANT	0.915	**0.929**	0.819	0.803	0.860
DOC	0.561	0.556	0.456	0.271	**0.574**
LOC	0.916	**0.918**	0.880	0.720	0.899
DATE	0.895	0.896	0.881	0.839	**0.908**
ORG	0.916	0.913	0.875	0.791	**0.918**
PERS	**0.968**	**0.968**	0.951	0.862	0.967
TIME	0.500	0.609	0.471	0.000	**0.700**
MON	0.955	**0.960**	0.906	0.915	0.919
MISC	0.344	**0.386**	0.249	0.138	0.359
ART	0.737	**0.759**	0.639	0.508	0.757
PCT	**1.000**	0.989	0.961	0.977	0.973
Overall	**0.890**	0.889	0.855	0.762	0.887

C. LLM Performance Across Evaluation Stages

Model	Bronze			Silver			Gold		
	Zero-Shot	Few-Shot	CoT	Zero-Shot	Few-Shot	CoT	Zero-Shot	Few-Shot	CoT
GPT-4o	**0.67**	**0.71**	**0.60**	**0.68**	**0.71**	**0.61**	**0.72**	**0.71**	**0.68**
Gemma-3-27B-IT	0.39	0.67	0.40	0.41	0.69	0.43	0.56	**0.71**	0.58
Gemma-2-27B-IT	0.45	0.62	0.38	0.49	0.66	0.40	0.58	0.70	0.51
Gemma-2-9B-IT	0.42	0.49	0.42	0.46	0.54	0.47	0.55	0.62	0.60
Phi-4	0.38	0.48	0.36	0.43	0.53	0.41	0.52	0.61	0.51
Qwen-2.5-14B-Instruct	0.42	0.50	0.36	0.44	0.53	0.38	0.53	0.57	0.48
Qwen-2.5-7B-Instruct	0.34	0.36	0.30	0.36	0.38	0.33	0.45	0.45	0.44
DeepSeek-R1-Distill-Qwen-14B	0.34	0.11	0.35	0.36	0.13	0.38	0.42	0.13	0.46
Gemma-2-2B-IT	0.16	0.30	0.25	0.20	0.37	0.28	0.28	0.47	0.36
Qwen-2.5-3B-Instruct	0.18	0.33	0.20	0.22	0.37	0.23	0.28	0.45	0.30
Llama-3.2-3B-Instruct	0.17	0.28	0.13	0.24	0.41	0.23	0.30	0.45	0.25
Phi-3-mini-4k-instruct	0.16	0.27	0.19	0.19	0.32	0.24	0.23	0.39	0.29
Llama-3.1-8B-Instruct	0.14	0.23	0.14	0.18	0.29	0.18	0.25	0.37	0.23
Aya-expanse-8b	0.23	0.03	0.23	0.31	0.03	0.28	0.34	0.03	0.29
Aya-101	-	0.31	-	-	0.38	-	-	0.41	-

D. Zero-Shot Performance of Generalist Models

Model	Bronze	Silver	Gold
gliner-multi-v2.1	**0.53**	**0.53**	**0.67**
gliner-multi-pii-v1	0.46	0.46	0.62
NuNER-Zero	0.41	0.41	0.58

E. Parameter Tuning with Different LoRA Parameters (80% Data)

Model	LoRA r=16	LoRA r=32	LoRA r=64
Qwen-2.5-14B-Instruct	0.851	0.851	**0.853**
Phi-4	0.869	0.871	**0.874**
Gemma-2-27B-IT	**0.865**	0.860	0.864
Gemma-3-27B-IT	0.867	0.879	**0.882**

F. Final SFT Results

Entity	**Qwen2.5-14B-Instruct**	**Phi-4**	**Gemma-2-27B-IT**	**Gemma-3-27B-IT**
JOB	0.624	0.638	**0.662**	0.642
PERIOD	0.667	0.714	0.742	**0.747**
QUANT	0.812	0.833	0.864	**0.897**
DOC	0.479	0.464	**0.537**	0.514
LOC	0.890	0.907	0.903	**0.929**
DATE	0.866	0.885	0.900	**0.906**
ORG	0.898	0.911	0.918	**0.923**
PERS	0.955	**0.967**	0.966	0.965
TIME	0.400	0.571	**0.824**	0.632
MON	0.950	0.958	**0.964**	0.953
MISC	**0.390**	0.314	0.311	0.350
ART	0.725	**0.774**	0.740	0.716
PCT	0.977	0.966	**0.994**	0.989
Overall	0.867	0.882	0.886	**0.888**

G. Comparison of Best-Performing Models Across Approaches

Entity	**Tuning**		**Prompting**		
	roberta-large-NER	**Gemma-3-27B-IT**	**GPT-4o**	**Gemma-3-27B-IT**	**GLiNER**
JOB	**0.699**	0.642	0.332	0.381	0.141
PERIOD	0.743	**0.747**	0.263	0.280	0.105
QUANT	**0.915**	0.897	0.475	0.000	0.155
DOC	**0.561**	0.514	0.122	0.000	0.111
LOC	0.916	**0.929**	0.775	0.782	0.705
DATE	0.895	**0.906**	0.650	0.738	0.663
ORG	0.916	**0.923**	0.809	0.757	0.672
PERS	**0.968**	0.965	0.900	0.870	0.863
TIME	0.500	**0.632**	0.308	0.111	0.154
MON	**0.955**	0.953	0.916	0.525	0.812
MISC	0.344	**0.350**	0.077	0.000	0.000
ART	**0.737**	0.716	0.289	0.000	0.175
PCT	**1.000**	0.989	0.910	0.949	0.867
Overall	**0.890**	0.888	0.724	0.713	0.669