SLENDER: Structured Outputs for SLM-based NER in Low-Resource Englishes

Nicole Ren*
GovTech Singapore
nicole_ren@tech.gov.sg

James Teo*
GovTech Singapore
james_teo@tech.gov.sg

Abstract

Named Entity Recognition (NER) for low-resource variants of English remains challenging, as most NER models are trained on datasets predominantly focused on American or British English. While recent work has shown that proprietary Large Language Models (LLMs) can perform NER effectively in low-resource settings through in-context learning, practical deployment is limited by their high computational costs and privacy concerns. Open-source Small Language Models (SLMs) offer promising alternatives, but the tendency of these Language Models (LM) to hallucinate poses challenges for production use. To address this, we introduce SLENDER, a novel output format for LM-based NER that achieves a three-fold reduction in inference time on average compared to JSON format, which is widely used for structured outputs. Our approach using Gemma-2-9B-it with the SLENDER output format and constrained decoding in zero-shot settings outperforms the en_core_web_trf model from SpaCy, an industry-standard NER tool, in all five regions of the Worldwide test set.

1 Introduction

Since the release of GPT-3 (Brown et al., 2020), Large Language Models (LLMs) have shown promising capabilities on Natural Language Processing (NLP) tasks (Wei et al., 2022). The direct use of closed-source LLMs such as ChatGPT for Named Entity Recognition (NER) has also been explored in zero-shot settings (Wei et al., 2023) and in specialised domains (Hu et al., 2024).

Although supervised models remain the predominant approach for NER, they face challenges in domains with scarce training data, such as low-resource settings (Wang et al., 2023) and cases with specialised label schemes such as in clinical domains (Hu et al., 2024). Fine-tuning of these models is possible but requires extensive labelled data which are scarce in low-resource settings.

Recent work has shown that proprietary LLMs can perform NER tasks effectively in low-resource settings through in-context learning (ICL) (Wang et al., 2023). However, their closed-source nature raises privacy concerns when processing sensitive data through third-party APIs. While open-source LLMs exist, the high compute costs of hosting them make them impractical for smaller organisations.

Open-source Small Language Models (SLMs) offer a viable alternative but come with their own challenges. Their tendency to hallucinate (Obaid ul Islam et al., 2025) poses difficulties for production use. To address this, we propose a strategy for Language Model (LM)-based NER tasks in low-resource Englishes that utilises a combination of: (i) a novel output format SLENDER, (ii) constrained decoding, and (iii) SLMs.

We conducted experiments on the Worldwide dataset (Shan et al., 2023) that contains low-resource Englishes from five geographical regions. Our approach using Gemma-2-9B-it with constrained decoding to output SLENDER format in zero-shot settings outperformed the en_core_web_trf model from SpaCy, an industry-standard NER tool (Honnibal et al., 2023), in F1 scores for all five regions of the Worldwide test set. Notably, SLENDER demonstrates a three-fold reduction in average inference time compared to JSON, which is widely used for structured outputs with LLMs. Our work makes the following contributions:

- We introduce SLENDER, a new and efficient output format for LMs that significantly reduces the number of tokens for structured output and inference time.
- We demonstrate that SLENDER coupled with constrained decoding in zero-shot settings enables Gemma-2-9B-it to outperform

*Equal contribution

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 836–849

the en_core_web_trf model from SpaCy, an industry-standard NER tool, in F1 scores for NER in low-resource Englishes. This eliminates two major barriers to NER applications in low-resource settings: the requirement for extensive labelled training data and the computational overhead of fine-tuning.

- We have refined the Worldwide test set with consistent annotations[1] to support future research in low-resource Englishes given the shortage of labelled datasets in this under-explored area.

2 Related Work

NER in Low-Resource Englishes. Despite the prevalence of English around the world, NER research has predominantly focused on American and British English variants, leaving a significant gap in understanding model performance for global English variants. Earlier work identified performance degradation for Western-English trained models in South African contexts (Louis et al., 2006).

Recent work by Shan et al. (2023) has also shown significant performance decrease when testing models trained on the CoNLL (Tjong Kim Sang and De Meulder, 2003) or OntoNotes (Weischedel et al., 2013) datasets on the global Worldwide dataset (Shan et al., 2023), but found minimal performance degradation with models trained on a combination of Worldwide with either CoNLL or OntoNotes.

NER Output Format. NER datasets often use the BIOES format to mark tokens with their entity class and position. Formats like BIOES have been found to be challenging for GPT-3 since they require each position in the input text to be aligned with each position of classes in the label sequence, leading to the novel use of special tokens such as "@@" and "##" to mark entities found within the text (Wang et al., 2023).

The consequence of using such a format is that the NER task for LLMs is limited only to a single entity type at a time. This constrains the practical application of LLMs for NER tasks, as real-world scenarios typically require the simultaneous identification of multiple entity types. Moreover, high token consumption for NER in these formats (Figure 1) can increase the time taken per task significantly.

[1]The dataset is available at https://github.com/njacl2025/slender-worldwide-dataset

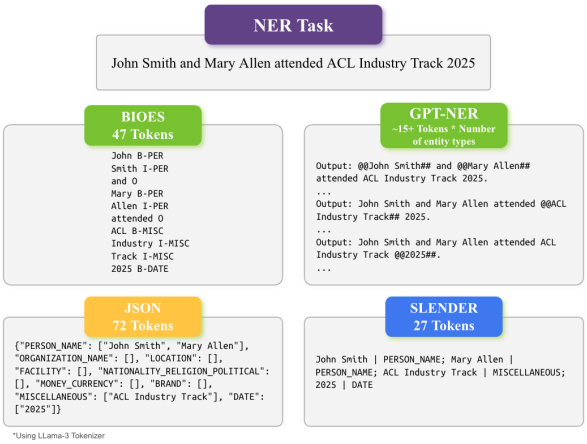

Figure 1: Comparison of token consumption of NER output formats. Tokens are counted using Llama-3 Tokenizer (Llama Team, AI @ Meta, 2024) as an example.

Our work contributes to this space by introducing SLENDER, a token-efficient output format for NER tasks using LMs that is capable of handling multiple entity types within the text simultaneously. SLENDER shows a significant reduction in the time taken for token generation compared to JSON.

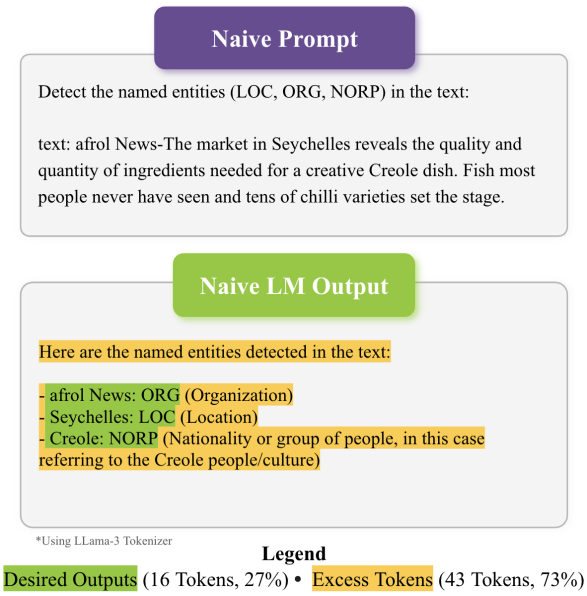

Figure 2: Token consumption for naïve LM output based on a naïve prompt. In this example, 71% of tokens are not required to complete the NER task. This example uses Llama-3.1-405B-Instruct for generation and Llama-3 Tokenizer (Llama Team, AI @ Meta, 2024) for counting the tokens.

3 Method

SLMs offer promising capabilities in NER, yet they present unique challenges. Despite their small size, SLMs still require substantial compute, which can

impede inference speed and diminish their viability for real-world applications. Moreover, smaller LMs tend to be more susceptible to hallucinations (Obaid ul Islam et al., 2025), potentially impacting their performance in NER tasks. To address these issues, we employ the following strategies:

Prompt Engineering. To increase the efficacy of SLMs in NER tasks, prompt engineering techniques like ICL provide additional context through task-specific demonstrations to prime the LM for NER tasks. The SLENDER output format complements this through its simplistic design to minimise the overhead in maintaining output structure unlike other complex structured formats.

This approach also avoids the typical behaviour of the LM to use a naïve output structure that tends to contain superfluous tokens. As seen in Figure 2, without this prevention, the LM outputs extra unnecessary tokens instead of completing the task efficiently.

Constrained Decoding. The use of structured output formats introduces innate rules that can be enforced during token generation. By applying constrained decoding, the likelihood of the model generating non-format conforming hallucinations can be significantly reduced (Geng et al., 2023).

3.1 SLENDER NER Output Format

SLENDER employs a linear representation methodology for entities extracted from the source text. In this approach, each entity is represented as a pair, comprising the entity itself and its corresponding entity type classification, with these elements being separated by a pipe symbol '|'. When multiple entities are present within the same source text, they are delimited using semicolons ';'. This minimalist syntactical approach demonstrates notable advantages over conventional JSON formats, particularly in terms of structural efficiency. The reduction in tokens required to maintain structural integrity results in significant speed gains for LM-based NER.

A trade-off is that it does not retain the positional information of the extracted entities, potentially making it more difficult to disambiguate identical entities appearing in different contexts within the same text. We also note that more tokens may be generated when entities of the same type appear more than once, as the entity type must be repeated for each instance. We observe that 43.7% of the Worldwide (Shan et al., 2023) test set has multiple entities of the same type.

3.2 In-context Learning (ICL)

0-shot, 3-shot and 5-shot ICL were tested for this study. The 3-shot and 5-shot implementations consist of one standard null example (text containing no entities), and K-1 randomly selected examples from the training set where K is the number of examples.

To ensure robust evaluation and to mitigate sampling variance, the few-shot trials were conducted using three random seeds for sampling examples from the training set and we reported the averaged performance on the test set. While bespoke, high-quality examples would be optimal for ICL, they are often impractical to obtain in real-world settings due to the vast diversity of scenarios. By selecting random examples from a standard dataset, our work provides a more realistic assessment of LM-based NER in practical settings.

3.3 NER Prompt Structure

We utilise the following prompt engineering techniques:

Model Priming. Our prompt gives the SLM a role as a "Named Entity Recognition System", incorporating clear tasks with label definitions and few-shot examples to guide the task execution. See Appendix A for the entity type definitions and Appendix B for the full prompt structure.

Pseudo XML. The prompt utilises Pseudo XML to organise content in a structured format and embed section-specific meta-information within the prompt.

Residual Bins for Entity Types. Additional entity types are weaved into the NER task to catch common false positives such as Food which were commonly misclassified as Miscellaneous.

3.4 Constrained Decoding

Constrained decoding is a technique to improve the validity of LM output formats by directing the LM generation process. The technique limits next-token predictions to only tokens that adhere to a predefined rule. In our study, constraints are applied to the SLM's at each generation step to enforce valid output structure that conform to JSON and SLENDER using the LMFE [2] and Guidance [3] constrained decoding libraries respectively.

[2] https://github.com/noamgat/lm-format-enforcer

[3] https://github.com/guidance-ai/guidance

4 Experiment

4.1 Dataset

The trials were conducted using the Worldwide dataset (Shan et al., 2023), which comprises English newswire articles from low-resource contexts including Asia, Africa, Latin America, the Middle East, and Indigenous Commonwealth (indigenous Oceania and Canada). We used the Stanza toolkit (Qi et al., 2020) to preprocess the dataset which contains 9 labels: Organization, Miscellaneous, Person, Money, Location, Facility, NORP (national, organizational, religious or political identity), Date and Product.

Improvements to Dataset Labels: To enhance annotation consistency with the dataset's published label definitions, we conducted a manual review of the annotations. This process revealed several inconsistencies. For instance, religious references such as "Allah", which is an Arabic word for God, were initially annotated as PERSON, despite the definitions excluding deities.

Similarly, event references such as "Covid-19" showed inconsistent labelling, appearing as both DATE and MISCELLANEOUS across different instances. We standardised such cases as MISCELLANEOUS to better reflect their semantic nature as events rather than temporal references. More examples can be found in Appendix C.

Given the scarcity of datasets for low-resource English NER, one of our key contributions is the release of this enhanced version of the Worldwide test set with refined annotations[4] to promote further research in this under-explored area.

4.2 Baseline

Model. We used the en_core_web_trf[5] transformer model from SpaCy, an industry-standard NER tool (Honnibal et al., 2023) as baseline. As this model, hereafter referred to as SpaCy, is trained on OntoNotes (Weischedel et al., 2013), we condense the labels into the Worldwide classes. See Appendix D for class mappings.

Output format. For a baseline output format, we used JSON, a common structured format that is widely used to obtain structured outputs from LMs. For the NER task, JSON organises entities hierarchically by entity classes, where each class serves as a key mapping to an array of corresponding entity mentions. To create a strong baseline, we applied prompt engineering to ensure valid JSON output by instructing the LM to include all 9 classes as keys to avoid hallucinations observed from having optional fields in preliminary experiments. See Appendix B for the full prompt format. We did not use the BIOES format as the baseline due to its documented challenges for LLMs (Wang et al., 2023), which are further exacerbated in SLMs.

4.3 Models

Microsoft (2024) popularised the term "Small Language Model" in the industry with their release of Phi-4, a 14-billion parameter SLM that surpasses much larger models on various benchmarks. In our experiments, we focus on instruction-tuned models under 10 billion parameters. This includes Meta's Llama-3-8B-Instruct (Llama Team, AI @ Meta, 2024), Microsoft's Phi-3.5-mini-Instruct (Microsoft Research, 2024) and Google's Gemma-2-9B-it (Gemma Team, Google DeepMind, 2024).

Post-Training-Quantisation is a widely adopted strategy for reducing the computational demand of a LM by decreasing the precision of model weights, albeit at the cost of model degradation. Research indicates that higher quantisation levels generally preserve model performance (Li et al., 2024). We chose the GGUF Q5_K_M quantisation scheme as a reasonable balance between model compression and performance retention.

4.4 SLM Inference

In each NER task, the SLM performs multi-class NER across all 9 entity classes defined in Worldwide test set simultaneously. For few-shot experiments, we ensure regional relevance by constructing prompts with randomly selected examples from the corresponding region's training set. When reporting the overall scores across regions or entities, we compute the micro-averaged F1 score to account for the variation in frequency of different classes across regions in the dataset (Shan et al., 2023).

4.5 Results

4.5.1 F1 Score Comparisons Across Regions

Gemma-2-9B-it with SLENDER and constrained decoding in a zero-shot setting outperforms the baseline, SpaCy, across all regions of the Worldwide test set (Table 1). The performance advantage of SLENDER is notable for Africa and Asia,

[4]The dataset is available at https://github.com/njacl2025/slender-worldwide-dataset

[5]https://huggingface.co/spacy/en_core_web_trf

Model	Format	Constrain	K-shot	Africa	Asia	IDG	Latam	ME
SpaCy (Baseline)	—	—	—	73.46	75.03	64.18	69.53	69.35
JSON Output Format								
Phi-3.5-mini	JSON	No	5-shot	50.81	57.58	50.75	49.23	52.02
Llama-3.1-8B	JSON	Yes	0-shot	68.77	70.29	63.35	67.26	64.07
Gemma-2-9B	JSON	No	0-shot	71.48	71.43	**71.95**	**70.08**	**70.38**
Gemma-2-9B	JSON	No	3-shot	72.79	72.29	**73.56**	**70.26**	**71.43**
Gemma-2-9B	JSON	No	5-shot	72.46	71.55	**73.33**	68.15	**69.98**
Gemma-2-9B	JSON	Yes	0-shot	68.73	69.19	**68.91**	67.68	67.41
Gemma-2-9B	JSON	Yes	3-shot	71.34	70.87	**71.80**	68.97	69.06
Gemma-2-9B	JSON	Yes	5-shot	70.66	70.87	**70.68**	68.46	68.57
SLENDER Output Format								
Phi-3.5-mini	SLENDER	No	5-shot	46.53	48.50	48.45	46.05	47.00
Llama-3.1-8B	SLENDER	No	5-shot	60.72	69.19	60.48	61.74	59.38
Gemma-2-9B	SLENDER	No	0-shot	66.71	72.83	**66.34**	66.83	69.05
Gemma-2-9B	SLENDER	No	3-shot	72.50	**79.06**	**72.78**	**74.50**	**72.09**
Gemma-2-9B	SLENDER	No	5-shot	72.66	**77.77**	**72.92**	**74.94**	**72.78**
Gemma-2-9B	SLENDER	Yes	0-shot	**74.43**	**78.35**	**69.86**	**75.17**	**74.74**
Gemma-2-9B	SLENDER	Yes	3-shot	71.90	**77.14**	**72.13**	**72.28**	**70.33**
Gemma-2-9B	SLENDER	Yes	5-shot	72.01	**75.92**	**70.03**	**72.46**	**71.25**

Table 1: F1 scores on Worldwide test set. All SLMs are of the instruct variant, with names shortened for brevity. SLENDER surpassed (bold) SpaCy for all five regions and outperformed JSON by achieving the highest (underlined) F1 scores for four out of five regions. For both Africa and Asia, only SLENDER-based approaches successfully surpassed SpaCy's strong baseline. For brevity, best-performing configurations are shown (see Appendix E for full results). IDG and ME refers to Indigenous and Middle East respectively.

where only SLENDER-based approaches successfully surpassed SpaCy's strong baseline. Furthermore, SLENDER outperforms JSON by achieving the highest F1 scores in four out of five regions, demonstrating the significant advantages of using the SLENDER format for NER tasks across diverse geographical contexts.

4.5.2 F1 Score Comparisons Across Entities

Using Gemma-2-9B-it, the best-performing SLM in our trials (Table 1), SLENDER achieved superior performance on six out of the nine entity classes in the Worldwide test set (Figure 3). This success was distributed across both constrained and unconstrained implementations of SLENDER – constrained decoding excelled for Organization, Product and Miscellaneous while unconstrained decoding performed better for Location, Date and Facility.

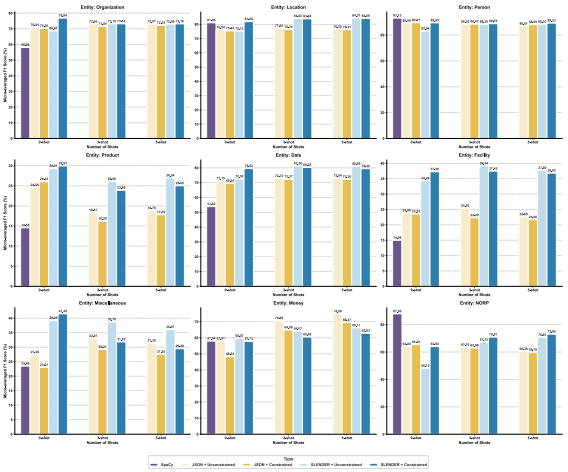

Figure 3: Entity-level F1 scores of Gemma-2-9B-it on Worldwide test set where SLENDER achieved superior performance on six out of nine entities.

In zero-shot settings, we observe that constrained decoding consistently improves F1 scores when using SLENDER. This is likely due to the novelty of the format for SLMs. However, this advantage diminishes in few-shot scenarios, suggesting that explicit demonstrations with SLENDER formats provided sufficient guidance for the SLM to maintain the SLENDER format with comparable F1 scores. Interestingly, constrained decoding shows minimal benefits for JSON and degrades performance in some cases. We hypothesise that this may be due to the widespread use of JSON and po-

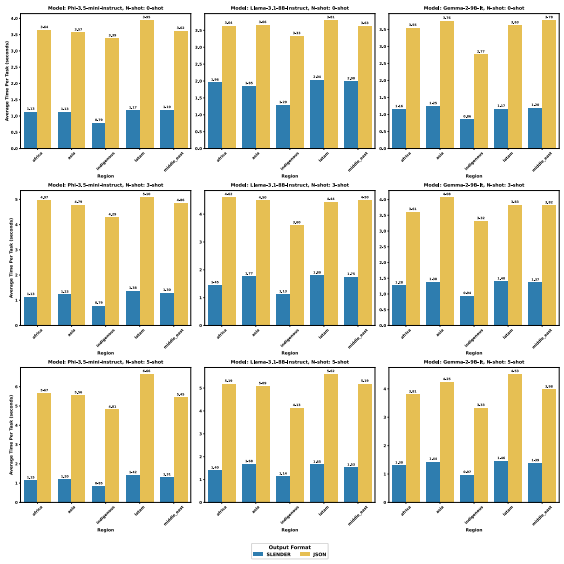

Figure 4: Average time taken per NER task on Worldwide test set. SLENDER strongly outperforms JSON with 3x reduction in time taken and tokens generated (see Appendix F).

tential prevalence in training data for SLMs, such that the use of constraints may force the model to deviate from its learnt generation patterns, resulting in suboptimal sequences.

SpaCy demonstrated notable strengths in specific entities, outperforming SLMs in Person and NORP. This strength likely stems from these entities' consistent representations across global English variants. For example, common NORP entities such as "Iranian", "Muslims", "Turkish" can be found across diverse regional datasets from Asia, Latin America and Middle East, suggesting that these entity types maintain consistent representations in their occurrences across English variants.

4.5.3 Efficiency Comparison

To evaluate format efficiency in isolation, we focused our efficiency analysis on non-constrained decoding scenarios, thereby eliminating potential confounding effects from implementation-specific overheads in constrained decoding libraries. Our evaluation demonstrates that the SLENDER format consistently achieves significant efficiency improvements over JSON in all configurations across models, regions, and K-shot examples. On average, SLENDER achieves a three-fold reduction in the average inference time (Figure 4). The efficiency gains of SLENDER have significant implications for real-world LM-based NER applications, where both processing speed and structured output formats are critical.

While JSON is a popular choice for structured outputs with LMs, its verbose syntax requirements involve a substantial number of structural tokens such as ' " ', '{', '}', '[', ']', which can significantly increase token count per query. We also acknowledge that there is room to improve the efficiency of complete JSON formats as the requirement for the keys to contain all entity classes can lead to extra tokens despite the absence of many entity classes in the input text. Nevertheless, this was necessary to create a strong baseline in F1 score performance (Figure 3) as it helps to address the observed tendency of SLMs to omit lower frequency labels. Future work can explore other methods to reduce the issues observed with optional fields within the JSON while improving token efficiency.

5 Conclusion

We introduced SLENDER, a novel output format for NER using LMs that demonstrates substantial advantages over the widely used JSON format. Our evaluation shows that SLENDER achieves a three-fold reduction in average inference time while improving F1 scores in challenging low-resource English contexts. The efficiency gains are especially valuable for real-world deployments to address critical concerns of latency and computational costs when using SLMs. The significant improvements of SLENDER highlights the importance of efficient output format design, an often overlooked avenue for optimising the performance of SLMs. As research continues to explore methods to make SLMs more practical for production use, our findings may have broader implications for other structured prediction tasks using LMs beyond NER.

6 Limitations

Dataset. Our work was evaluated using only the Worldwide dataset due to our focus on low-resource Englishes, an understudied area that has not been examined recently until Shan et al. (2023). For future work, we hope to evaluate with other datasets to understand the performance of SLENDER in other low-resource settings such as low-resource non-English languages. We also did not encounter any edge cases impacted by the use of the reserved tokens ';' and '|' in SLENDER. We plan to explore the robustness of SLENDER in future work.

In-context Learning Examples. Our current

implementation retrieves examples through random selection from the train set. This provides a realistic assessment reflecting real-world scenarios where curated examples are often impractical. We observed that only 8.10% of our 864 few-shot experimental results had zero variation in F1 scores among the 3 seeds used for random sampling. The largest delta observed was a drop of 36.25%, even after excluding cases of entities with counts less than 30 in its specific region. While we reported the average F1 score to reduce variability, future work can explore different retrieval methods such as kNN-based retrieval using entity-level representations (Wang et al., 2023) to retrieve demonstrations that are semantically close to the input text.

Constrained Decoding Libraries. We observed difficulties with using Guidance for constrained trials on SLENDER using Llama-3-8b and Phi-3.5-mini due to its engine migration during our research period. To preserve analytical integrity, affected trials were not included in our primary findings but are documented in Appendix E. Future work can compare the use of other constrained decoding libraries and models.

Limited Compute. Due to resource constraints, all our experiments were conducted on NVIDIA T4 (2018) GPUs, which offer substantially lower computational capability compared to newer GPUs. This restricted our choice of models in trials, which we hope to expand on in future work.

References

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, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Gemma Team, Google DeepMind. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952, Singapore. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2023. explosion/spaCy: v3.7.2: Fixes for APIs and requirements.

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.

Shiyao Li, Xuefei Ning, Luning Wang, Tengxuan Liu, Xiangsheng Shi, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. 2024. Evaluating quantized large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Llama Team, AI @ Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Anita Louis, Alta de Waal, and Cobus Venter. 2006. Named entity recognition in a south african context. In *Proceedings of the 2006 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries*, SAICSIT '06, pages 170–179, Somerset West, South Africa. South African Institute for Computer Scientists.

Microsoft. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Microsoft Research. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Saad Obaid ul Islam, Anne Lauscher, and Goran Glavaš. 2025. How much do llms hallucinate across languages? on multilingual estimation of llm hallucination in the wild. *Preprint*, arXiv:2502.12769.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alexander Shan, John Bauer, Riley Carlson, and Christopher Manning. 2023. Do "English" named entity recognizers work well on global englishes? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11778–11791, Singapore. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *Preprint*, arXiv:2304.10428.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with chatgpt. *ArXiv*, abs/2302.10205.

Ralph Weischedel and 1 others. 2013. OntoNotes release 5.0. Web Download. LDC Catalog No.: LDC2013T19.

A LM NER Task Entity Definitions

LM Label	Original Label	Definition
PERSON_NAME	PERSON	Full or partial names of specific individuals, including first names, last names, middle names, and initials (e.g., 'John Smith', 'J.K. Rowling'). Exclude titles (Dr., Mr.), relationship terms (mother, boss), and possessive forms.
LOCATION	LOCATION	Named geographical entities including countries, cities, states, streets, addresses, landmarks, mountains, rivers, oceans, continents and other physical places (e.g., 'France', 'Mount Everest', '123 Main Street').
ORGANIZATION	ORGANIZATION	Named entities that represent groups of people working together for a purpose, including companies, government agencies, non-profits, schools, sports teams, and political parties (e.g., 'Apple Inc.', 'United Nations', 'Manchester United').
NATIONALITY_-RELIGION_-POLITICAL	NORP	Terms referring to national, ethnic, religious, political identities, or ancestry/heritage (e.g., 'American', 'Buddhist', 'Republican', 'Hispanic', 'Celtic', 'Anglo-Saxon'). Include demonyms, adjectives describing these identities, and terms referring to historical or cultural lineage.
DATE	DATE	Temporal references to specific calendar dates, including full dates, partial dates, named days, holidays, and time periods (e.g., 'January 15, 2023', 'last Tuesday', 'Christmas', 'summer of 2020').
MISCELLANEOUS	MISCELLANEOUS	Other named entities that don't fit in above categories, such as events (e.g., 'World Cup'), awards (e.g., 'Nobel Prize'), works of art/media (e.g., 'Mona Lisa', 'Star Wars'), and other proper nouns.
MONEY_-CURRENCY	MONEY	Monetary values and currency names, including specific amounts with currency indicators, currency symbols, and names of currencies (e.g., '$100', 'Euro', '5 million dollars').
FACILITY	FACILITY	Named physical structures or installations with specific purposes, including buildings, stadiums, airports, bridges, and monuments (e.g., 'Empire State Building', 'JFK Airport', 'Golden Gate Bridge')
BRAND	PRODUCT	Names of commercial products, services, and their associated brands or trademark names (e.g., 'iPhone', 'Coca-Cola', 'Nike Air Max'). Do not include the generic product type unless it's part of the branded name.

Table 2: Mapping of Entity Class Names for Worldwide Dataset to Labels used in LM Prompts with definitions and examples.

B Prompt Structure

Prompt

```
{SYSTEM_PROMPT_START_TOKEN}You are a state-of-the-art Named Entity Recognition
system.{SYSTEM_PROMPT_END_TOKEN}{USER_PROMPT_START_TOKEN}Your task is to detect
all variations of Entities in the given Text according to the Definitions. The
same Entities may appear in different spellings and lengths. Detect all unique
variants of the Entities within the provided Text. Do not repeat yourself.

{OUTPUT_FORMAT_INSTURCTIONS}

Here are the Definitions of the Entities to detect:

{DEFINITIONS}

Here are some examples:

{EXAMPLES}

Now detect the Entities in the following Text:

<text>{INPUT_TEXT}</text>{USER_PROMPT_END_TOKEN}{ASSISTANT_START_TOKEN}<entities>
{MODEL STARTS GENERATING}
```

NER Task-specific Inputs

```
{OUTPUT_FORMAT_INSTURCTIONS}:
Output the Entities detected using the
following format ```<entities> ENTITY_1 |
CATEGORY_1; ENTITY_2 | CATEGORY_2; ... ;
ENTITY_N | CATEGORY_N </entities>```. If
there are no Entities found, respond with
```<entities>NIL</entities>```.

{DEFINITIONS}:
<definition>PERSON_NAME: Full or partial
names ... </definition> ...

<definition>{Residual Bin Entities E.g.
Food}: ...</definition> ...

{EXAMPLES}:
<text>I love ...</text>
<entities>NIL</entities>

<text>Nigel Milne ...</text>
<entities>Nigel Milne | PERSON_NAME;
Ron...</entities> ...
```

**Model-specific Format Tokens (Where Applicable)**

```
E.g. Gemma-2-9b-it Prompt format:
<bos><start_of_turn>user
<end_of_turn>
<start_of_turn>model
```

Figure 5: Prompt Structure for SLENDER Output Format.

.

Prompt

```
{SYSTEM_PROMPT_START_TOKEN}You are a state-of-the-art Named Entity Recognition
system.{SYSTEM_PROMPT_END_TOKEN}{USER_PROMPT_START_TOKEN}Your task is to detect
all variations of Entities in the given Text according to the Definitions. The
same Entities may appear in different spellings and lengths. Detect all unique
variants of the Entities within the provided Text. Do not repeat yourself.

Output a valid JSON object with keys: {ENTITY_TYPES}. Each key maps to a list of
unique entities of that type found in the text. If there are no entities found,
respond with an empty {EMPTY_RESULTS_EXAMPLE}.

Here are the Definitions of the Entities to detect:

{DEFINITIONS}

Here are some examples:

{EXAMPLES}

Now detect the Entities in the following Text:
<text>
{INPUT_TEXT}</text>{USER_PROMPT_END_TOKEN}{ASSISTANT_START_TOKEN}<entities>{MODEL
STARTS GENERATING}
```

NER Task-specific Inputs

```
{ENTITY_TYPES}:
PERSON_NAME, ORGANIZATION_NAME, LOCATION,
FACILITY, DATE, MONEY_CURRENCY, BRAND,
NATIONALITY_RELIGION_POLITICAL, MISCELLANEOUS, <
Residual Bin Entities>

{EMPTY_RESULTS_EXAMPLE}
JSON ```{}```


{DEFINITIONS}:
<definition>PERSON_NAME: Full or partial names
... </definition> ...

<definition>{Residual Bin Entities E.g. Food}:
...</definition> ...

{EXAMPLES}:
<text>I love ...</text>
<entities>NIL</entities>

<text>Nigel Milne ...</text>
<entities>Nigel Milne | PERSON_NAME;
Ron...</entities> ...
```

Model-specific Format Tokens (Where Applicable)

```
E.g. Gemma-2-9b-it Prompt format:
<bos><start_of_turn>user
<end_of_turn>
<start_of_turn>model
```

Figure 6: Prompt Structure for JSON Output Format.

.
```

C Worldwide NER Label Improvements

Text	Original Class Label	Improved Class Label	Rationale
Officials said the two started firing from a rooftop but were "quickly eliminated by mujahideen with the help of Allah the almighty".	Allah PERSON	Allah O	"Allah", which is the Arabic word for God, is re-labelled as O (not an entity). This is due to the exclusion of deities from PERSON according to the dataset documentation.
Xiaomi's decision to tap Vietnam as its latest production base drew public attention as it followed similar moves by major global smartphone makers to move parts of their supply chain from China to Southeast Asia in search of lower costs and more stable production output during Covid-19.	Covid-19 DATE	Covid-19 MISCELLANEOUS	"Covid-19" in this context refers to the pandemic as an event or phenomenon, therefore falling under MISCELLANEOUS.
But it was not so easy for me to manage when I encountered Germans. Antisemitism typified the Germans even in those days, and the toxic hatred of Jews welled up in them already then."	Antisemitism PERSON	Antisemitism O	"Antisemitism" describes a form of prejudice, rather than a name of humans. As it does not fall into any of the other classes, it is re-labelled as O (not an entity).
First, Shaked noticeably refrained from mentioning whether she would join a government led by opposition leader Benjamin Netanyahu. She mentioned Netanyahu's name only once in her speech: "The housing crisis and high cost of living are not interested in 'yes Bibi, no Bibi.'"	Bibi O	Bibi PERSON	"Bibi" in this context refers to the nickname of Benjamin Netanyahu, and there is a clear connection between "Netanyahu" and "Bibi" in the same text hence re-labelled as PERSON.
Other projects included the Electric Company buildings, Haifa's central train station and the old building in the northern city's Bnei Zion Medical Center.	Haifa MISCELLANEOUS	Haifa LOCATION	"Haifa" in this context refers to the city in Israel, and is therefore re-labelled as LOCATION instead of MISCELLANEOUS.
The first democratically-elected President of South Africa, and the country's first Black leader, died in December 2013 at age 95.	December O	December DATE	"December" refers to the month and thus labelled as DATE.
On Friday morning, Syrian media said that Israel had hit Damascus, killing three military forces and injuring seven more.	Friday MISCELLANEOUS	Friday DATE	"Friday" refers to the day and thus labelled as DATE.

Table 3: Examples of improvements to labels and corresponding rationale for the Worldwide dataset.

D SpaCy Mappings

SpaCy Label	Worldwide Label
PERSON	PERSON
ORG	ORGANIZATION
GPE	LOCATION
LOC	LOCATION
FAC	FACILITY
DATE	DATE
TIME	DATE
NORP	NORP
LANGUAGE	NORP
MONEY	MONEY
PRODUCT	PRODUCT
EVENT	MISCELLANEOUS
WORK_OF_ART	MISCELLANEOUS
LAW	MISCELLANEOUS
PERCENT	DROP
QUANTITY	DROP
ORDINAL	DROP
CARDINAL	DROP

Table 4: SpaCy Label to Worldwide Label Mapping

E F1 Scores on Worldwide Test Set for All Experiments

Model	Format	Constrain	K-shot	Africa	Asia	IDG	Latam	ME
SpaCy (Baseline)	—	—	—	73.46	75.03	64.18	69.53	69.35
JSON Output Format								
Phi-3.5-mini	JSON	No	0-shot	50.43	52.98	49.17	52.48	46.56
Phi-3.5-mini	JSON	No	3-shot	50.36	57.17	50.50	49.17	51.39
Phi-3.5-mini	JSON	No	5-shot	50.81	57.58	50.75	49.23	52.02
Llama-3.1-8B	JSON	No	0-shot	68.68	69.65	62.24	66.27	63.15
Llama-3.1-8B	JSON	No	3-shot	43.12	55.19	43.09	43.06	39.27
Llama-3.1-8B	JSON	No	5-shot	43.16	59.12	43.38	56.72	46.66
Gemma-2-9B	JSON	No	0-shot	71.48	71.43	**71.95**	**70.08**	**70.38**
Gemma-2-9B	JSON	No	3-shot	72.79	72.29	**73.56**	**70.26**	**71.43**
Gemma-2-9B	JSON	No	5-shot	72.46	71.55	**73.33**	68.15	**69.98**
Phi-3.5-mini	JSON	Yes	0-shot	49.31	52.77	49.41	50.46	45.71
Phi-3.5-mini	JSON	Yes	3-shot	48.74	55.22	49.79	47.55	50.81
Phi-3.5-mini	JSON	Yes	5-shot	48.32	54.79	49.19	47.46	50.78
Llama-3.1-8B	JSON	Yes	0-shot	68.77	70.29	63.35	67.26	64.07
Llama-3.1-8B	JSON	Yes	3-shot	64.24	68.83	59.38	62.31	62.67
Llama-3.1-8B	JSON	Yes	5-shot	64.76	70.00	58.69	64.14	63.82
Gemma-2-9B	JSON	Yes	0-shot	68.73	69.19	**68.91**	67.68	67.41
Gemma-2-9B	JSON	Yes	3-shot	71.34	70.87	**71.80**	68.97	69.06
Gemma-2-9B	JSON	Yes	5-shot	70.66	70.87	**70.68**	68.46	68.57
SLENDER Output Format								
Phi-3.5-mini	SLENDER	No	0-shot	45.38	48.73	51.55	45.37	43.13
Phi-3.5-mini	SLENDER	No	3-shot	45.69	46.55	49.23	41.43	43.53
Phi-3.5-mini	SLENDER	No	5-shot	46.53	48.50	48.45	46.05	47.00
Llama-3.1-8B	SLENDER	No	0-shot	52.84	57.09	49.37	51.86	51.00
Llama-3.1-8B	SLENDER	No	3-shot	59.33	66.78	58.21	59.60	56.59
Llama-3.1-8B	SLENDER	No	5-shot	60.72	69.19	60.48	61.74	59.38
Gemma-2-9B	SLENDER	No	0-shot	66.71	72.83	**66.34**	66.83	69.05
Gemma-2-9B	SLENDER	No	3-shot	72.50	**79.06**	**72.78**	**74.50**	**72.09**
Gemma-2-9B	SLENDER	No	5-shot	72.66	**77.77**	**72.92**	**74.94**	**72.78**
Phi-3.5-mini	SLENDER	Yes	0-shot	29.21	31.49	29.60	28.67	27.27
Phi-3.5-mini	SLENDER	Yes	3-shot	29.01	32.92	27.45	27.98	27.58
Phi-3.5-mini	SLENDER	Yes	5-shot	29.65	32.14	26.89	30.09	29.24
Llama-3.1-8B	SLENDER	Yes	0-shot	47.83	52.29	41.68	47.46	44.85
Llama-3.1-8B	SLENDER	Yes	3-shot	48.58	53.69	43.12	47.50	45.95
Llama-3.1-8B	SLENDER	Yes	5-shot	49.95	54.94	44.48	51.49	47.42
Gemma-2-9B	SLENDER	Yes	0-shot	**74.43**	**78.35**	**69.86**	**75.17**	**74.74**
Gemma-2-9B	SLENDER	Yes	3-shot	71.90	**77.14**	**72.13**	**72.28**	**70.33**
Gemma-2-9B	SLENDER	Yes	5-shot	72.01	**75.92**	**70.03**	**72.46**	**71.25**

Table 5: F1 scores on the Worldwide test set for all experiments conducted. All SLMs in the table are of the instruct variant, with names shortened for simplicity. IDG and ME refers to Indigenous and Middle East respectively. Experiments outperforming the SpaCy baseline are bolded and best-performing ones in each region are underlined.

F Average Tokens Generated per NER Task

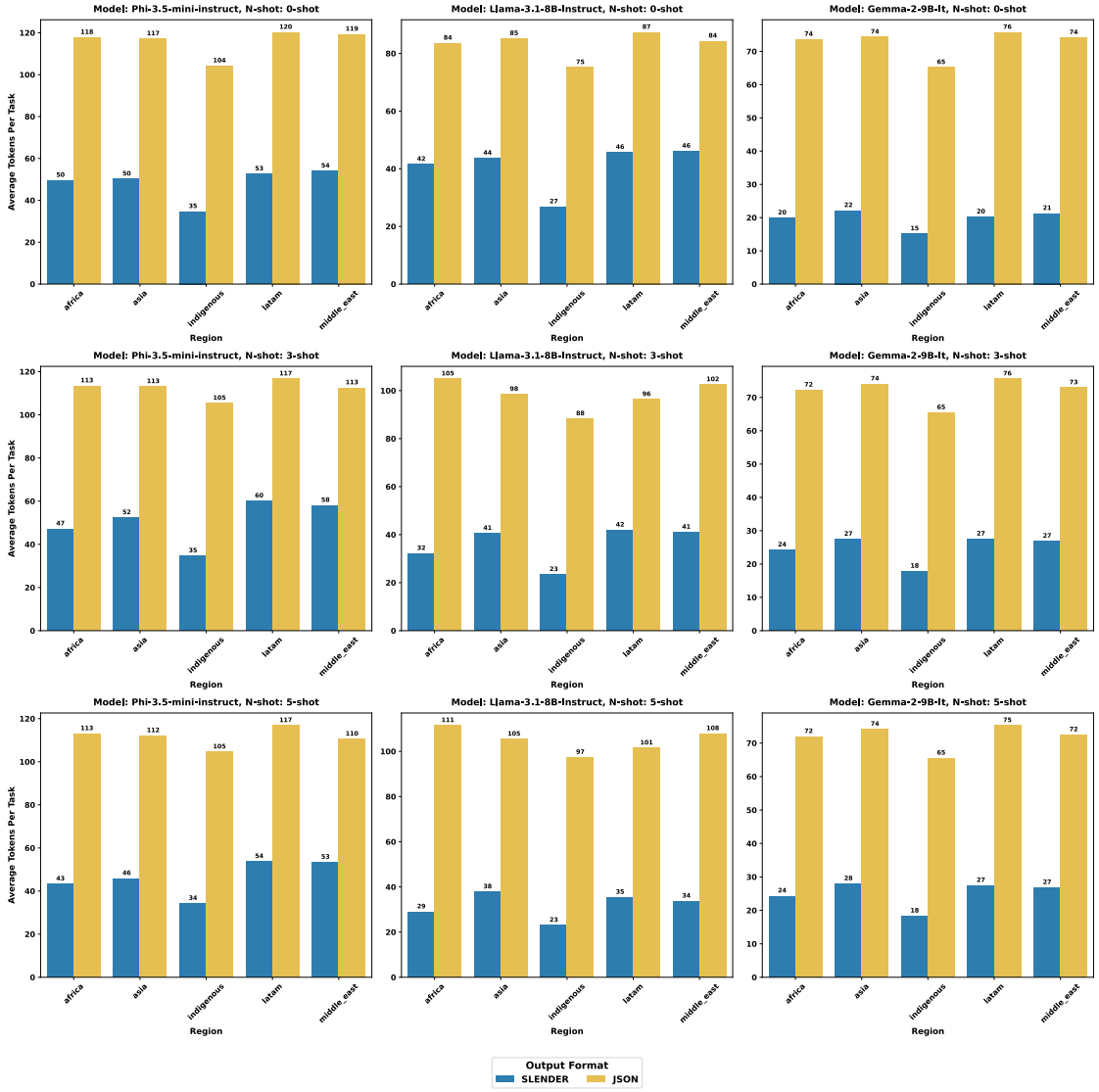

Figure 7: Average tokens generated per NER task on the Worldwide test set. SLENDER strongly outperforms JSON format, with on average threefold reduction in tokens generated.