Robust and Efficient Indonesian Span-Based Named Entity Recognition via Compact GLiNER

Towards Enhanced Retrieval-Augmented Generation

Mukhlish Fuadi
Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
7022231007@student.its.ac.id

Adhi Dharma Wibawa
Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia | Department of Medical Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
ad_wibawa@its.ac.id (corresponding author)

Surya Sumpeno
Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia | Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
surya@te.its.ac.id

Received: 2 March 2026 | Revised: 18 April 2026 | Accepted: 1 May 2026

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | *DOI: https://doi.org/10.48084/etasr.18482*

ABSTRACT

Efficient Named Entity Recognition (NER) integration is crucial for improving retrieval precision and fact verification in Retrieval-Augmented Generation (RAG) systems. However, the development of Indonesian NER still faces challenges, including dataset heterogeneity, inefficient tokenization, and trade-offs between NER performance and inference efficiency. Although span-based frameworks such as Generalist and Lightweight Named Entity Recognition (GLiNER) offer greater flexibility than conventional Beginning-Inside-Outside (BIO) approaches, available GLiNER models are generally multilingual and have not been optimized for Indonesian characteristics. This research proposes a compact GLiNER model specifically for Indonesian, utilizing a pruned mDeBERTa with a 30k vocabulary as the encoder backbone. We built a large-scale Indonesian GLiNER training corpus by combining heterogeneous NER datasets and augmenting them with controlled translations, resulting in 56,210 training, 6,707 validation, and 9,411 test samples. Experimental results show that the proposed model achieves an F1 score of 76.58%, surpassing the IndoBERT-based GLiNER baseline (74.70%) at max_len = 192 tokens, with consistent improvements on major entities. Deployment-oriented evaluation shows up to 8 × CPU-based inference acceleration (482 ms vs. 3,897 ms per sample). In long-context evaluation (max_len = 384), the proposed model outperforms multilingual GLiNER by more than 12 F1 points while maintaining a significantly lower memory footprint on both GPU and CPU. These advantages validate the model's potential as a reliable metadata filter component for RAG architectures.

Keywords-Named Entity Recognition (NER); GLiNER; span-based NER; Retrieval-Augmented Generation (RAG); Transformer models; efficient NLP

I. INTRODUCTION

In the current Generative Artificial Intelligence (AI) landscape, Named Entity Recognition (NER) integration is increasingly vital, particularly as a supporting component of the Retrieval-Augmented Generation (RAG) architecture [1, 2]. Named entities function as metadata filters that narrow the vector search space (pre-retrieval) and as fact verification anchors (post-generation) to mitigate model hallucinations [3]. However, real-time implementation requires a balance between extraction performance and computational efficiency that conventional architectures struggle to achieve, especially given constraints on document heterogeneity and local infrastructure limitations [4]. Furthermore, since RAG systems typically process multi-sentence document chunks, there is a critical

demand for NER models with long-context robustness to ensure entity coherence across extended sequences.

Recent developments in NER have been dominated by Transformer architectures, which have significantly improved performance [5]. However, most systems remain limited to rigid token-level Beginning-Inside-Outside (BIO) schemes that are prone to propagation errors in long or nested entities, which are common in legal documents and corporate reports [6]. As an alternative, the span-based NER approach models entities directly as token ranges [7], significantly reducing the problem of cascading label misclassification caused by initial prediction errors in the BIO scheme [8]. In addition to providing richer range representations and robustness to Out-of-Vocabulary (OOV) conditions [9], this approach offers greater structural flexibility, better aligned with the dynamic information-extraction needs of RAG systems [10].

Generalist and Lightweight Named Entity Recognition (GLiNER) was introduced as a span-based NER framework that offers flexible label prediction via a marker-based approach [11]. This approach allows for a clearer separation between text representation and label definition. Despite operating with significantly lower parameter sizes and computational costs, GLiNER outperformed generative models such as ChatGPT in several tests. Although more efficient than Large Language Models (LLMs), currently available GLiNER models are dominated by multilingual variants that have not been optimized for Indonesian's morphological characteristics. The use of massive multilingual vocabularies often leads to representation inefficiencies due to excessive subword segmentation (over-fragmentation), which directly increases computational load and creates latency bottlenecks when integrated into complex inference pipelines or local deployment scenarios [12]. This challenge is exacerbated by the heterogeneity of local NER datasets and the scarcity of local models that can handle long document contexts stably [13, 14], creating an urgent need for compact yet robust models.

To address these challenges, this study proposes GLiNER-ID, an Indonesian span-based NER model optimized for deployment efficiency and high context stability. Our approach integrates three main strategies: (1) construction of a large-scale training corpus through harmonization of heterogeneous datasets and augmentation of translation data into a consistent span-based format; (2) utilization of a compact language-oriented encoder (vocabulary pruned) as a backbone to maximize representation efficiency without changing the core architecture; and (3) experimental evaluation that focuses not only on performance, but also on robustness to input length variation and practical inference efficiency.

The main contributions of this research are summarized as follows:

- Standardization and compilation of a large-scale Indonesian GLiNER corpus from heterogeneous sources.
- Development of a compact Indonesian GLiNER model using vocabulary-pruned mDeBERTa.
- Comprehensive evaluation of performance, efficiency, and long-context robustness for RAG suitability.
- Empirical validation of the model's superiority over IndoBERT and multilingual baselines.
- Public release of the model via a public repository to support reproducibility [15].

II. METHODOLOGY

A. Datasets

1) Indonesian Native Named Entity Recognition Corpora

To train the GLiNER model focused on Indonesian, this study combines several Indonesian NER datasets from various domains and sources. This combination aims to cover variations in language style, topics, and annotation schemes that reflect the actual use of Indonesian.

The Indonesian NER datasets used include idner-news [16], NERGrit [14], NERGrit Corpus [17], NER UI [18], NER UGM [19], and NERP [20]. A summary of the composition of each dataset is shown in Table I, including the division of training, validation, and testing data, as well as the original annotation scheme for each dataset.

TABLE I. COMPOSITION OF INDONESIAN NATIVE NER DATASETS

Dataset	Train	Dev	Test	Original entity tags
idner-news [16]	1,464	367	509	PER, ORG, LOC
NER UI [18]	7,654	850	2,126	PERSON, ORGANIZATION, LOCATION
NER UGM [19]	8,437	935	2,343	PERSON, ORGANIZATION, LOCATION, QUANTITY, TIME
NERGrit [14]	1,672	209	209	PERSON, ORGANISATION, PLACE
NERGrit Corpus [17]	12,514	2,520	2,397	CRD, DAT, EVT, FAC, GPE, LAW, LOC, MON, NOR, ORD, ORG, PER, PRC, PRD, QTY, REG, TIM, WOA, LAN
NERP [20]	6,720	840	840	PPL, PLC, EVT, FNB, IND
Total	38,461	5,721	8,424	

All datasets initially differed in their annotation formats, including variations in the number of columns, separators, and label conventions, and thus a normalization process was required before they could be integrated within the GLiNER framework. All BIO annotations were processed to extract basic entity labels. Labels with the same semantic meaning but different names were grouped into one consistent entity category (e.g., PER, PERSON, and PPL were mapped to Person). In addition to general entities, rich entity types available in certain datasets, such as Date, Money, Law, Facility, Product, Language, and Work of Art, were retained as long as they could be mapped explicitly and consistently. IND labels in the NERP dataset that do not represent valid named entities were treated as Outside (O) labels and removed from the annotation.

After label normalization, all BIO annotations were converted to the span-based format used by GLiNER, where each entity is represented as a pair of start–end token indices

along with its entity type. This process also included annotation validation and cleaning to ensure span index consistency and training data quality.

2) Translated Named Entity Recognition Data

In addition to the original Indonesian NER dataset, this study also utilizes additional NER data obtained through a controlled translation process from the English-language dataset pile-mistral-v0.1 [21]. This dataset was chosen because it provides NER annotations with thousands of entity types in a span-based format compatible with the GLiNER framework, and it exhibits high diversity in contexts and sentence structures.

The translation process was carried out using Google Neural Machine Translation (GNMT) services accessed programmatically. Given the large scale of the dataset, translation execution was implemented iteratively using a fault-tolerant checkpointing mechanism. This mechanism ensures process stability by periodically saving progress and enabling automatic recovery (resume capability) without data loss in the event of network interruptions.

To ensure annotation integrity during translation, this study implements a masking-based, entity-preservation mechanism. As illustrated in Figure 1, this approach separates the translation process of sentence context and named entities. In the initial stage, each entity span in the source sentence is substituted with a unique placeholder token (e.g., ENT 0) to isolate the syntactic structure from the entity terminology. Translation is then performed separately on the masked text and the mapped entity list. In the reconstruction stage, placeholders in the translated sentence are replaced with target entities, and the span index is automatically recalibrated. This method ensures that entity boundaries in the Indonesian output remain accurate despite changes in word structure or phrase order due to the translation process. Qualitative observations across multiple samples confirm that this method can produce contextually representative target texts without distorting the original meaning, making it suitable for model training. After translation, the dataset is divided into training, development, and test subsets at 90%:5%:5%. The division is done randomly with a fixed random seed to ensure reproducibility. With this composition, 17,749, 986, and 987 are obtained for the train, dev, and test sets, respectively.

Fig. 1. Entity-preserving translation pipeline.

3) Final Dataset Statistics

The final dataset used in this study is a combination of the original Indonesian NER dataset and the translated dataset. The combination was done separately for each split (train, dev, and test) to maintain consistent data separation. All data were then randomized using a fixed random seed to ensure an even distribution of sentences from various sources during training.

Overall, the final dataset consists of 56,210 training sentences, 6,707 validation sentences, and 9,411 test sentences. This dataset covers various entity types, ranging from general entities such as Person, Location, and Organization, to rich entities such as Date, Money, Law, Product, Actor, and Sports Team. The distribution of entity labels is shown in Figure 2, which displays the 19 entity types with the highest frequencies, with the remaining entity types grouped into the OTHER category. This documented conversion process and the public model [15] ensure full reproducibility for community-led data preparation.

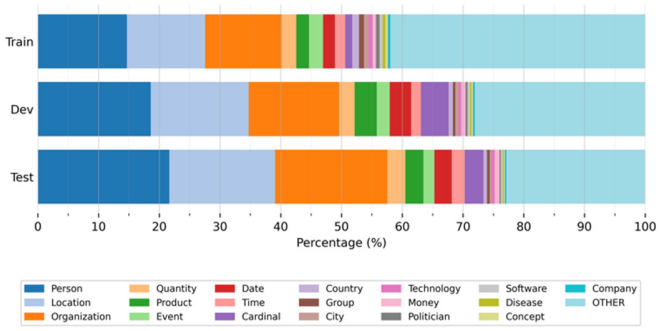

Fig. 2. NER label distribution.

B. Model Architecture

1) Generalist and Lightweight Named Entity Recognition Framework

This study uses GLiNER as the main framework for performing span-based NER. Unlike the BIO token-based classification approach, GLiNER formulates NER as a structured span-prediction task using a marker-based approach [11]. The input text is processed as a sequence of tokens, and the model directly predicts the start and end token indices that form an entity, along with its entity type.

2) Backbone Encoders

To evaluate the influence of encoder selection on the performance and efficiency of GLiNER in Indonesian, this study compares several Transformer-based backbone encoders. All encoders are integrated into the GLiNER framework with identical architectural configurations, allowing performance differences to be directly attributed to the encoders' characteristics.

a) GLiNER-ID

The proposed model utilizes the mdeberta-hybrid-30k [22] encoder as its backbone. This encoder is a compact variant of mDeBERTa-v3-base [23] developed through language-aware frequency-based vocabulary pruning [22]. This technique reduced the original 250k vocabulary to 30k, optimized for Indonesian (70%) and English (30%) to enhance computational efficiency while maintaining performance. In this study, we adopt this model as a pretrained backbone without internal

architectural modifications, aligning with the base architecture of gliner_multi-v2.1 [11].

b) IndoBERT (Baseline)

As the main baseline, this study uses IndoBERT-base-p2 as the encoder backbone in GLiNER. IndoBERT is one of the most widely used pretrained Indonesian language models in Natural Language Processing (NLP) research [14]. This model was selected to provide a fair and neutral comparison, given the similarity of the Transformer architecture and its widespread use as a baseline in Indonesian NER studies.

c) Multilingual GLiNER

As an additional comparison, this study also evaluates the multilingual GLiNER model, gliner_multi-v2.1 [11]. This model is used only in the testing stage (test set) to provide a reference point regarding the performance of the proposed model compared to common multilingual solutions, rather than as the main baseline.

3) Model Configuration

To ensure fair comparability, all models were trained using identical hyperparameter configurations. A standard marker-based scheme (markerV0) was adopted with a hidden size of 768, dropout of 0.4, and a maximum span length limit of 12 tokens. The fine-tuning process was carried out for 10,000 optimization steps with a batch size of 16. A differential learning rate strategy was applied to maintain the stability of the backbone representation, with a learning rate of 1e-5 for the encoder and 5e-5 for other task-specific components. All training used a cosine learning rate scheduler with a warm-up ratio of 0.1 and weight decay of 0.01 to prevent overfitting. These settings ensure that the observed performance variations objectively reflect the intrinsic characteristics of the encoder backbone, while isolating them from training configuration inconsistencies.

C. Evaluation Protocol

NER performance was evaluated via exact span-matching in a closed-set scenario, using micro-averaged F1 score to handle unbalanced label distributions. To ensure a fair comparison across different tokenization schemes, an exact text-span-matching mechanism was implemented. Validation is based on the similarity of entity text strings rather than token indices, mitigating bias from tokenization shifts. The evaluation procedure also includes applying a safe decoder patch to ensure inference stability and granular analysis of the ten entity classes with the highest frequency. Furthermore, to ensure statistical reliability, each experimental configuration was validated through multiple runs. The performance metrics reported in the subsequent sections represent stable, representative values observed across these independent iterations, confirming the consistency of the results.

Deployment efficiency is evaluated using an oracle-label inference approach to estimate the upper bound of computational performance. The metrics measured include average latency, throughput, and peak memory usage in GPU and CPU environments. All efficiency measurements were performed on a standardized test data subset that included a warm-up phase, with detailed hardware specifications recorded to ensure reproducibility and transparency of analysis on limited resources.

III. RESULTS AND DISCUSSION

A. Named Entity Recognition Performance

Before reviewing the aggregate performance metrics, a qualitative evaluation confirmed that both model variants have adequate semantic capabilities for identifying Indonesian-language entity structures. As shown in Figure 3, inference examples on the test set (max_len = 192) indicate that the model accurately predicts span boundaries and entity labels according to the ground truth, even in sentences with high entity density and various types, such as Organization, Product, Time, and Money. The success of predictions on this sample indicates that, fundamentally, both backbones are capable of adapting the span-based GLiNER mechanism. However, a more comprehensive quantitative analysis in Table II reveals significant differences in performance stability between the two.

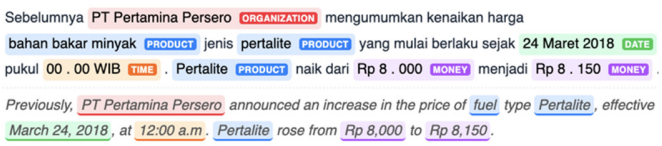

Fig. 3. Visualization of NER prediction (with English translation).

Table II summarizes the model performance using micro-averaged span-level F1 score in the closed-set scenario. In general, GLiNER-ID shows superior performance and greater stability than the baseline IndoBERT across all configurations. At short sequence lengths (max_len = 128), GLiNER-ID achieved an F1 score of 78.72%, slightly exceeding IndoBERT (78.00%). The architectural advantage of GLiNER-ID becomes more significant as the context length increases. At the same time, whereas IndoBERT experiences a sharp degradation to 74.70% at max_len = 192, GLiNER-ID maintains a much gentler decline, reaching 76.58%.

TABLE II. OVERALL NER PERFORMANCE

Model	Max_len	Precision (%)	Recall (%)	F1 score (%)
GLiNER + IndoBERT	128	84.54	72.41	78.00
	160	84.42	69.33	76.13
	192	84.05	67.23	74.70
GLiNER-ID	128	81.27	76.33	78.72
	160	81.02	74.44	77.59
	192	80.71	72.84	76.58
	384	79.69	71.26	75.24

In addition to its stability in medium contexts, GLiNER-ID demonstrates robustness that the baseline lacks, operating at max_len = 384 and achieving an F1 score of 75.24% without additional modifications. This score remains competitive compared to IndoBERT on much shorter sequences (max_len = 192). Component-level analysis reveals that IndoBERT tends to prioritize precision over recall, particularly on long sequences. In contrast, GLiNER-ID offers a better precision–recall balance, with higher recall. These findings confirm that

using the pruned mDeBERTa backbone in GLiNER provides improved robustness against input-length variations, effectively preserving long entity spans, which is a critical advantage over fixed sentence splitting, as it risks truncating vital context or entity boundaries in real-world deployment.

B. Impact of Maximum Sequence Length

A sensitivity analysis of the maximum sequence length (Figure 4) reveals differences in the degradation characteristics of performance between the two models. IndoBERT experienced a sharp decline from 78.00% (max_len = 128) to 74.70% (max_len = 192), indicating that adding context destabilizes the precision–recall balance. In contrast, GLiNER-ID shows a much more gradual decline (78.72% to 76.58%) and maintains competitive performance at 75.24% even at max_len = 384.

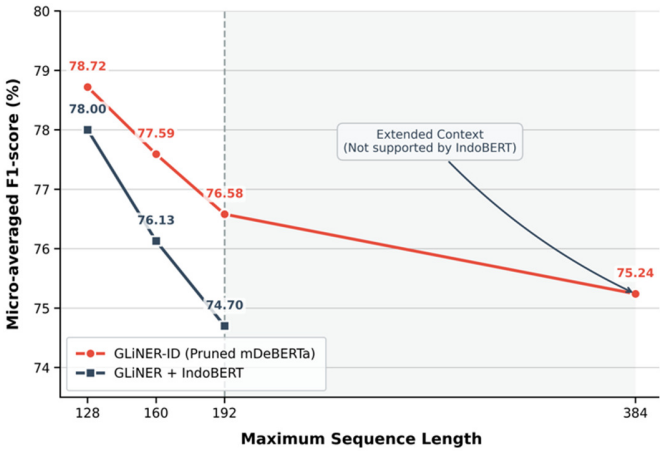

Fig. 4. Effect of maximum sequence length on NER performance.

Beyond performance metrics, architectural stability limits are a crucial differentiator. IndoBERT experiences runtime failures at configurations above max_len = 192, an inherent limitation due to the use of absolute positional embeddings and rigid fixed token type embeddings [14]. On the other hand, GLiNER-ID operates stably up to max_len = 384 thanks to the disentangled attention mechanism in the mDeBERTa backbone, which separates position and content representations [23]. This capability to handle long contexts provides significant practical advantages for Indonesian, where semantic units are often represented by long sequences of phrases, opening up opportunities for context exploration that the IndoBERT baseline cannot accommodate.

C. Per-Entity Performance Analysis

A granular analysis of the ten highest-frequency entities at max_len = 192 in Table III reveals that GLiNER-ID consistently outperforms the baseline in the entity categories with clear lexical boundaries. This model achieved superior F1 scores for Person (92.74% vs. 91.36%) and improvements of +1.05 and +0.75 points for Organization and Location, respectively. The most significant advantage is observed for the Money entity (97.12% vs. 95.24%), which was supported by better generalization of contextual-numerical patterns.

TABLE III. PER-ENTITY PERFORMANCE COMPARISON

Entity type	IndoBERT F1 score (%)	GLiNER-ID F1 score (%)
Person	91.36	92.74
Organization	84.04	85.09
Location	87.00	87.75
Cardinal	94.47	92.15
Product	73.08	72.83
Quantity	74.53	70.66
Date	91.69	92.12
Time	88.06	86.08
Event	55.58	57.11
Money	95.24	97.12

However, IndoBERT maintains its advantage in purely numerical categories such as Cardinal (94.47% vs. 92.15%) and Quantity, indicating a minor trade-off due to vocabulary simplification in highly explicit lexical patterns. For entities with high semantic ambiguity, such as Event, both models show lower performance, but GLiNER-ID still leads with 57.11% compared to 55.58%. Overall, these results confirm that compact encoders provide crucial selective improvements on dominant entities without sacrificing substantial competitiveness in minor categories.

D. Inference Efficiency

Efficiency evaluation (Tables IV and V) was conducted on a computing environment comprising an NVIDIA A100-SXM4 GPU (40GB) and an Intel Xeon Gold 5218 CPU (32 cores, 2.30 GHz). Testing used an oracle-label inference scheme to estimate the upper bound on performance. All experiments were run on a Linux environment with the PyTorch 2.5.1 framework and CUDA 12.4.

TABLE IV. INFERENCE EFFICIENCY BENCHMARK

Metric	Unit	GLiNER + IndoBERT	GLiNER-ID
CPU performance			
Avg latency	ms	3,896.98	482.24
Throughput	samples/s	0.26	2.07
Peak RAM	MB	1,450.52	1,452.99
GPU performance			
Avg latency	ms	10.29	18.96
Throughput	samples/s	96.88	52.65
Peak memory	MB	1,650.44	1,629.33

The inference benchmark in Table IV uses max_len = 192. In the GPU environment, the IndoBERT baseline recorded superior throughput (96.88 samples/s) and latency (10.29 ms) compared to GLiNER-ID (52.65 samples/s; 18.96 ms). Although peak memory consumption is relatively similar (~1.6 GB), this latency disparity is due to the complexity of tensor operations in the mDeBERTa disentangled attention mechanism, which places a greater burden on GPU parallel execution than the standard BERT architecture.

The opposite pattern was observed in CPU evaluation, where GLiNER-ID recorded up to 8× inference acceleration compared to IndoBERT (482 ms vs. 3.9 s per sample). Given that the sizes of both models are identical (~510 MB), this performance surge is not due to model reduction but rather to improved representation efficiency. Drastic vocabulary pruning significantly reduces the computational load of embedding

projections and matrix operations in the early layers, which are major CPU bottlenecks, effectively compensating for the complexity of mDeBERTa's attention. These findings confirm GLiNER-ID's suitability for real-world deployment on standard CPU-based infrastructure without GPU acceleration.

E. Comparison with Multilingual Generalist and Lightweight Named Entity Recognition

A comparative study with gliner_multi-v2.1 was conducted on a max_len = 384 configuration to validate the effectiveness of the language-oriented approach. The test results shown in Table V indicate that GLiNER-ID outperforms the multilingual model in both performance and resource efficiency. In a CPU environment, the proposed model achieved an F1 score of 75.28%, significantly outperforming (+12 points) the multilingual variant (63.36%). This advantage is reinforced by substantial memory efficiency; GLiNER-ID requires only 8.3 GB of system RAM, reducing memory requirements by up to 40% compared to multilingual models (13.3 GB) burdened by massive parameter sizes.

TABLE V. COMPARISON WITH MULTILINGUAL GLINER

Metric	Unit	GLiNER-ID	GLiNER-Multi-v2.1
Model properties			
Model size	MB	508.75	1,102.26
CPU performance			
F1 score	%	75.28	63.36
Latency	ms	519.89	464.50
Throughput	samples/s	1.92	2.15
Peak memory	MB	8,336.09	13,290.24
GPU performance			
F1 score	%	75.20	63.12
Latency	ms	19.50	18.90
Throughput	samples/s	51.22	52.84
Peak memory	MB	627.49	1,205.43

A similar pattern is seen in the GPU scenario, where GLiNER-ID maintains performance superiority (F1 score 75.20% vs. 63.12%) with much more efficient VRAM consumption (0.63 GB vs. 1.2 GB), even though the latency and throughput of both models are relatively identical. These findings confirm that hardware acceleration cannot fully mitigate the parameter inefficiencies of multilingual models. Instead, the vocabulary-pruned encoder approach offers the best cost-performance ratio, delivering high performance with a minimal memory footprint, ideal for production deployment.

F. Discussion

This experiment reveals a non-linear dynamic in which directed vocabulary pruning has been shown to exponentially improve CPU efficiency without degrading semantic capabilities. GLiNER-ID consistently outperforms the IndoBERT baseline at practical sequence lengths (max_len ≤ 192), driven by greater prediction stability on key entities (Person, Location, Organization), indicating alignment between the encoder vocabulary distribution and the morphological characteristics of Indonesian. Sensitivity analysis shows that increasing context length does not guarantee improved performance, as information saturation occurs at medium lengths. However, GLiNER-ID's technical advantage lies in its architectural stability up to max_len = 384, a capability that the baseline lacks, providing greater flexibility for handling long documents.

As external validation against community benchmark standards, we conducted a comparative evaluation with IndoBERT [13] using the public NER UI and NER UGM datasets. Recognizing the fundamental differences between span-based and token-based architectures, a rigorous evaluation protocol was applied to ensure comparability. Model span predictions were mapped to the IOB2 format using exact boundary matching and evaluated with the standard seqeval library, following the IndoLEM benchmark protocol. The experimental results confirm the superiority of the proposed model, achieving micro F1 scores of 91.77% on NER UI and 76.85% on NER UGM, surpassing those of IndoBERT (90.1% and 74.9%, respectively). These findings indicate that the span-based approach with an efficient backbone is highly robust in capturing named entities, even when evaluated using strict token-based metrics.

Beyond sentence-level evaluation, model utility is validated in long-document processing scenarios. Given the 384-token architectural limit, a boundary-aware sliding window inference mechanism is implemented. This mechanism splits the text into overlapping chunks while preserving sentence boundaries, using score-based deduplication to resolve redundancy in overlapping regions. Practical testing shows that this approach enables GLiNER-ID to extract entities from large-scale documents in their entirety, with precision controlled via threshold adjustment, making it a viable component for semantic enrichment in RAG systems.

Regarding efficiency, the performance disparity between GPUs and CPUs underscores the crucial role of compact encoder design. While GPU latency is dominated by attention complexity, CPU inference on GLiNER-ID achieves ~8× acceleration due to reduced embedding projection load. These findings, combined with comparisons against multilingual GLiNER, reinforce the argument for language specialization: language-oriented approaches not only reduce memory and parameter overhead but also yield higher performance than generalist models. However, the findings are limited to closed-set scenarios and oracle-label evaluations, which represent an upper bound of theoretical performance. The study focuses on monolingual optimization and does not aim to replace the cross-lingual capabilities of multilingual models.

IV. CONCLUSION

This study addresses a critical research gap in Indonesian Named Entity Recognition (NER) by investigating the trade-off between extraction performance and inference efficiency, particularly for Retrieval-Augmented Generation (RAG) architectures. Generalist and Lightweight Named Entity Recognition (GLiNER)-ID is developed by integrating a standardized multi-source corpus with a compact pruned-vocabulary encoder. Empirical evaluation shows that this approach consistently outperforms the IndoBERT baseline, achieving F1 scores of 78.72% (max_len = 128) and 76.58% (max_len = 192), while maintaining operational stability in long contexts (384 tokens), which are essential for document

processing in RAG scenarios, where the baseline fails. This performance consistency is further validated on standard community benchmarks, reinforcing the model's generalization capability.

The primary novelty of this work lies in adapting the span-based GLiNER framework with a compact, language-specialized backbone to optimize Indonesian entity extraction. The most distinctive advantage is observed in CPU inference scenarios, where GLiNER-ID achieves up to 8× acceleration compared to IndoBERT (482 ms vs. 3.9 s per sample). Furthermore, comparative analysis shows that the proposed model outperforms multilingual GLiNER by more than 12 F1 points in accuracy while reducing memory consumption by up to 40%. These findings confirm that language specialization with a compact encoder provides an effective trade-off between performance and computational cost, positioning GLiNER-ID as a practical solution for Indonesian NER in real-world production environments and paving the way for more precise and resource-efficient RAG systems.

DECLARATION OF COMPETING INTERESTS

The authors declare no conflicts of interest and have no financial interest to report.

ACKNOWLEDGMENT

This work was supported by the Indonesia Endowment Fund for Education (LPDP), Ministry of Finance, Republic of Indonesia, under Grant LOG-9715/LPDP/LPDP.3/2024. The computations described in this paper were performed using the Mahameru High Performance Computing Facility of the National Research and Innovation Agency (BRIN), Republic of Indonesia.

DATA AVAILABILITY

All datasets used in this study are publicly available and properly cited within the manuscript, including idner-news [16], NERGrit [14], NERGrit Corpus [17], NER UI [18], NER UGM [19], and NERP [20].

REFERENCES

[1] Z. Wang, H. Chen, G. Xu, and M. Ren, "A novel large-language-model-driven framework for named entity recognition," *Information Processing & Management*, vol. 62, no. 3, May 2025, Art. no. 104054, https://doi.org/10.1016/j.ipm.2024.104054.

[2] Y. Huang and J. X. Huang, "A Survey on Retrieval-Augmented Text Generation for Large Language Models," *ACM Computing Surveys*, Apr. 2026, https://doi.org/10.1145/3805774.

[3] T. Fan, J. Wang, X. Ren, and C. Huang, "MiniRAG: Towards Extremely Simple Retrieval-Augmented Generation." arXiv, Jan. 26, 2025, https://doi.org/10.48550/arXiv.2501.06713.

[4] I. Budi and R. R. Suryono, "Application of named entity recognition method for Indonesian datasets: a review," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 2, pp. 969–978, Apr. 2023, https://doi.org/10.11591/eei.v12i2.4529.

[5] N. Kaur, A. Saha, M. Swami, M. Singh, and R. Dalal, "Bert-Ner: A Transformer-Based Approach For Named Entity Recognition," in *2024 15th International Conference on Computing Communication and Networking Technologies*, Kamand, India, 2024, pp. 1–7, https://doi.org/10.1109/ICCCNT61001.2024.10724703.

[6] T. E. Moussaoui, C. Loqman, and J. Boumhidi, "Exploring the Impact of Annotation Schemes on Arabic Named Entity Recognition across General and Specific Domains," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 21918–21924, Apr. 2025, https://doi.org/10.48084/etasr.10205.

[7] W. L. Seow, I. Chaturvedi, A. Hogarth, R. Mao, and E. Cambria, "A review of named entity recognition: from learning methods to modelling paradigms and tasks," *Artificial Intelligence Review*, vol. 58, no. 10, July 2025, Art. no. 315, https://doi.org/10.1007/s10462-025-11321-8.

[8] J. Yu, B. Ji, S. Li, J. Ma, H. Liu, and H. Xu, "S-NER: A Concise and Efficient Span-Based Model for Named Entity Recognition," *Sensors*, vol. 22, no. 8, Apr. 2022, Art. no. 2852, https://doi.org/10.3390/s22082852.

[9] J. Fu, X. Huang, and P. Liu, "SpanNER: Named Entity Re-/Recognition as Span Prediction," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online, 2021, pp. 7183–7195, https://doi.org/10.18653/v1/2021.acl-long.558.

[10] U. Zaratiana, N. Tomeh, P. Holat, and T. Charnois, "Named Entity Recognition as Structured Span Prediction," in *Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures*, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 1–10, https://doi.org/10.18653/v1/2022.umios-1.1.

[11] U. Zaratiana, N. Tomeh, P. Holat, and T. Charnois, "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)*, Mexico City, Mexico, 2024, pp. 5364–5376, https://doi.org/10.18653/v1/2024.naacl-long.300.

[12] M. Fuadi, A. D. Wibawa, and S. Sumpeno, "Adaptation of Multilingual T5 Transformer for Indonesian Language," in *2023 IEEE 9th Information Technology International Seminar*, Batu Malang, Indonesia, 2023, pp. 1–6, https://doi.org/10.1109/ITIS59651.2023.10420049.

[13] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), 2020, pp. 757–770, https://doi.org/10.18653/v1/2020.coling-main.66.

[14] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China, 2020, pp. 843–857, https://doi.org/10.18653/v1/2020.aacl-main.85.

[15] "muchad/gliner-id." Hugging Face. https://huggingface.co/muchad/gliner-id.

[16] S. O. Khairunnisa, A. Imankulova, and M. Komachi, "Towards a Standardized Dataset on Indonesian Named Entity Recognition," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, Suzhou, China, 2020, pp. 64–71, https://doi.org/10.18653/v1/2020.aacl-srw.10.

[17] H. Lovenia *et al.*, "SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, FL, USA, 2024, pp. 5155–5203, https://doi.org/10.18653/v1/2024.emnlp-main.296.

[18] Y. Gultom and W. C. Wibowo, "Automatic open domain information extraction from Indonesian text," in *2017 International Workshop on Big Data and Information Security*, Jakarta, Indonesia, 2017, pp. 23–30, https://doi.org/10.1109/IWBIS.2017.8275098.

[19] M. Fachri, "Named entity recognition for Indonesian text using hidden Markov model," B.S. thesis, Universitas Gadjah Mada, Yogyakarta, Indonesia, 2014.

[20] D. Hoesen and A. Purwarianti, "Investigating Bi-LSTM and CRF with POS Tag Embedding for Indonesian Named Entity Tagger," in *2018 International Conference on Asian Language Processing*, Bandung,

Indonesia, 2018, pp. 35–38, https://doi.org/10.1109/IALP.2018.8629158.

[21] "urchade/pile-mistral-v0.1." Datasets at Hugging Face, Aug. 21, 2024. [Online]. Available: https://huggingface.co/datasets/urchade/pile-mistral-v0.1.

[22] M. Fuadi, A. D. Wibawa, and S. Sumpeno, "Efficient Transformer Models via Language-Aware Frequency-Based Vocabulary Pruning," *IEEE Access*, vol. 14, pp. 50993–51006, 2026, https://doi.org/10.1109/ACCESS.2026.3679735.

[23] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing." arXiv, Nov. 18, 2021, https://doi.org/10.48550/arXiv.2111.09543.