GenoVarDis@IberLEF2024: Automatic Genomic Variants and Related Diseases using Named Entity Recognition with Large Language Models

Víctor Manuel Oliveros[1]

[1]Universidad de Granada, Escuela Técnica Superior de Ingeniería Informática y Telecomunicaciones (ETSIIT), Calle Periodista Daniel Saucedo Aranda, s/n, 18014 Granada, Spain

Abstract

This paper presents a student's proposal from the University of Granada for the GenoVarDis shared-task at IberLEF2024, focusing on Automatic Disease and Procedure Coding using Named Entity Recognition (NER). We utilize the GenoVarDis shared-task dataset, comprising scientific literature in Spanish on genomic variants, genes, and related diseases. We propose three language models for our task: (i) GPT, despite not being specifically trained for NER, we evaluate its performance in this domain, (ii) RoBERTa, pretrained in the biomedical domain with Spanish texts by the Plan de Tecnologías del Lenguaje (PlanTL) of the Government of Spain, originally designed for Fill-Mask tasks, has been fine-tuned for Named Entity Recognition (NER), and (iii) GLiNER, a compact NER model that excels in identifying arbitrary entity types, outperforming both ChatGPT and fine-tuned Large Language Models (LLMs) in zero-shot evaluations on various NER benchmarks. According to official results, we ranked second overall.

Keywords

Named Entity Recognition, Token classification, Spanish, Biomedical text processing, Biomedical NER, GLiNER

1. Introduction

Within the realm of healthcare, the automation of disease is paramount for enhancing efficiency and patient outcomes. When fused with state-of-the-art Large Language Models (LLMs), Named Entity Recognition (NER) techniques emerge as a powerful toolset for driving this coding automation. These advanced techniques represent a significant step towards revolutionizing healthcare processes, ensuring precision and effectiveness in automated coding tasks [1].

The motivation behind the application of NLP models for automatic disease coding lies in the quest to enhance the medical sector's ability to provide reliable, precise, and high-quality diagnoses for patients. By harnessing evolving algorithms and cutting-edge technologies, such as GPT and BERT, we aim not only to streamline the diagnostic and treatment process but also to minimize errors and optimize medical record management. This approach not only modernizes the medical field but also aligns it with current technological advancements, ensuring safer, more reliable, and efficient services for patients [2].

Named Entity Recognition (NER) is not a new topic in natural language processing (NLP), including language identification, code-switching, sentiment analysis, or machine translation. The goal of the GenoVarDis: NER in Genomic Variants and related Diseases for the Biomedical NLP shared task [3] at IberLEF 2024 [4] is to facilitate the automatic identification of genomic variants and associated diseases within biomedical text data, thereby enabling advancements in the field of Named Entity Recognition (NER) specifically tailored to the Spanish biomedical domain. Therefore, the proposed task is as follows: to perform Named Entity Recognition (NER) to automatically identify and classify entities related to genomic variants and diseases within Spanish biomedical text data.

The primary challenge of this task lies in the limited resources available for Named Entity Recognition (NER) and genomic variants in Spanish. It utilizes a distinctive corpus that includes a wide range of

IberLEF 2024, September 2024, Valladolid, Spain

voliverosvillena@gmail.com (V. M. Oliveros)
https://www.linkedin.com/in/victormov/ (V. M. Oliveros)
0009-0004-0968-678X (V. M. Oliveros)

mutations and entities related to variants, such as genes, diseases, and symptoms in Spanish. These resources are primarily sourced through translations from English and carefully curated by human experts. This proposal is notable for its innovation and importance, as it addresses significant obstacles in identifying variant-related entities [3].

Our approach We leverage three state-of-the-art models for Named Entity Recognition (NER) tasks in the biomedical domain: GPT-3.5 Turbo, RoBERTa, and GLiNER. Each model offers unique strengths that contribute to the overall effectiveness of our approach. GPT-3.5 Turbo, known for its efficiency and versatility, is employed for iterative analysis tasks due to its rapid processing capabilities. Specifically, RoBERTa, available on Hugging Face (https://huggingface.co/) under the model name PlanTL-GOB-ES/roberta-base-biomedical-es, has been pretrained using a byte version of Byte-Pair Encoding (BPE) with a vocabulary size of 52,000 tokens, following the approach employed in the original RoBERTa model. The pretraining involves a masked language model training at the subword level, utilizing the same hyperparameters as the RoBERTa base model. Additionally, GLiNER, a recent advancement in entity identification and classification [5], provides additional precision and performance improvements, particularly in challenging scenarios. GLiNER, available on Hugging Face under the model name urchade/gliner_medium-v2.1 (Medium variant with 209M parameters), has obtained very good results with zero-shot scenarios. This comparative analysis aims to shed light on whether recent advancements in models like GLiNER or GPT have the potential to yield notable results in this domain.

Additionally, our experiments were conducted in the Google Colab environment, specifically the free version, which posed certain limitations in terms of memory allocation, both in RAM and GPU resources.

2. Overview of the shared task

The shared task is centered around the analysis of biomedical text data for Automatic Disease and Procedure Coding using Named Entity Recognition (NER) in Spanish.

For the Named Entity Recognition (NER) task, we adopted the BIO schema for RoBERTa, where tokens are categorized into entities with labels indicating their position within the entity. However, for GLiNER, which offers flexibility in entity labeling, we maintained the same set of labels as in RoBERTa, but without the strict adherence to the BIO schema. GPT-3.5 Turbo, on the other hand, directly identifies entities within the clinical text data without the need for explicit labeling. The types of entities addressed in our task include: DNAMutation, SNP, DNAAllele, NucleotideChange/BaseChange, OtherMutation, Gene, Disease and Transcript.

Metrics The metrics employed for evaluation are precision, recall, and F1-score to assess the performance of our NER models. Additionally, we will address the limitations of metrics like accuracy for this task and provide further insights into their applicability. The criterion used for considering a named entity valid is exact match.

Data We only used the dataset provided by the competition to train our models. The dataset consists of six files in TSV (Tab-Separated Values) format, specifically containing information from tmVar3 documents from PubMed along with another 136 cases from PubMed in Spanish and from SciELO.

The six files have been partitioned into pairs to serve the purposes of training, validation, and testing, adhering to a split ratio of 70%-10%-20%. Each pair comprises a "text" file, featuring three columns: pmid, filename, and text (representing the biomedical case text), and an "annotation" file containing comprehensive annotations detailing: pmid, filename, mark, label, offset1, offset2, and span. These annotations precisely delineate the entity's position within the text of its associated document.

The training set consists of 427 clinical cases, while the validation set comprises 70 cases, and the test set contains 136 cases.

3. Our models

We briefly introduce the models used for Named Entity Recognition (NER) in the biomedical domain in this section. Firstly, the GPT model employed is GPT-3.5 Turbo, from the OpenAI API, for its efficiency and versatility in iterative analysis tasks. Known for rapid processing capabilities, this model allows us to perform tasks such as summarization, translation, and entity recognition directly from clinical text data [6]. Unlike traditional NER models that require explicit labeling, GPT-3.5 Turbo identifies entities based on prompts provided for each clinical case. This model is based on the transformer architecture and uses unidirectional masked self-attention. With up to 175 billion parameters, GPT-3.5 Turbo captures complex language nuances, making it suitable for generating precise and relevant responses in the biomedical domain.

Moving on to RoBERTa, this model is a biomedical pretrained language model which was previously fine-tuned specifically for Spanish biomedical tasks [7]. Its training corpus comprises a diverse collection of biomedical texts in Spanish sourced from various reputable sources such as Scielo and PubMed. This collection includes clinical cases, scientific articles, research papers, and other forms of medical literature. The model's training involved masked language model training at the subword level, following a similar approach to the original RoBERTa model. Notably, RoBERTa's training corpus comprises approximately 963 million tokens sourced from various biomedical sources, ensuring its robustness and effectiveness in biomedical text analysis tasks. Its primary function is "Fill Mask," where it predicts missing words or phrases in a given sentence. However, we will fine-tune it for Named Entity Recognition (NER) tasks.

Introduced in November 2023, GLiNER stands out as a specialized model tailored for Named Entity Recognition (NER). Its architecture, centered around a Bidirectional Language Modeling Transformer, enables precise identification and classification of entities, showcasing remarkable performance, especially in zero-shot scenarios. Notably, GLiNER's parallel entity extraction capability sets it apart, allowing for efficient extraction of entities across multiple categories simultaneously. With variants spanning from small to large, GLiNER exhibits competitiveness across diverse domains. Despite its recent introduction leading to limited documentation and resources, it offers promising potential for real-world applications, particularly in the medical realm. We aim to fine-tune our dataset effectively by leveraging GLiNER's capabilities, accessible through its Hugging Face library.

3.1. Data Exploration

Data exploration delves into the intricacies of our dataset, providing a detailed insight into its composition, distribution, and fundamental characteristics. This stage is crucial in any analytical project as it lays the groundwork for understanding the nature of the data and guiding subsequent decisions in the modeling or analysis process.

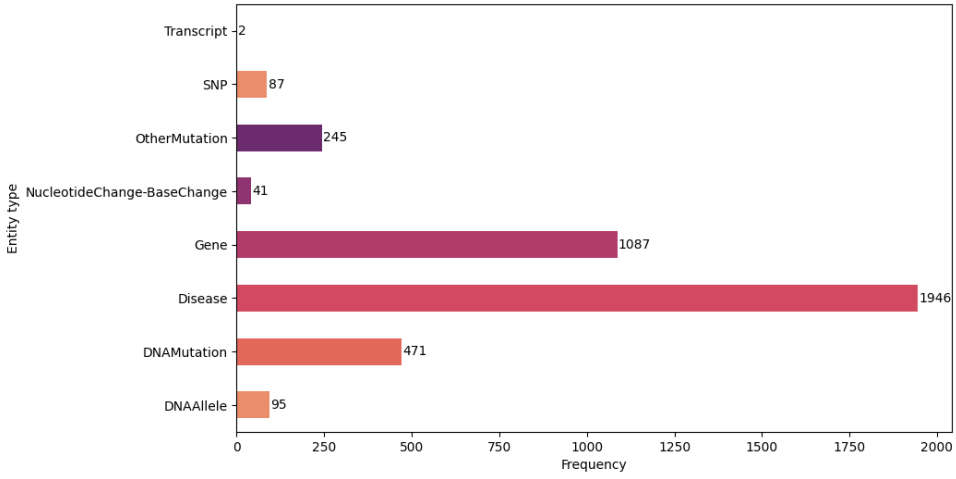

Figure 1: Unique entities distribution by type in the training and dev data.

Given the prevalence of diseases, genes, and genetic mutations in the dataset [Figure 1], it's evident that these entities hold significant importance in the biomedical texts under consideration. Consequently, in the development and fine-tuning of NER models, it's imperative to prioritize the accurate identification and classification of these entities. However, the relatively lower frequencies of nucleotide changes and transcripts suggest that these entities may pose greater challenges for the models to learn effectively.

Furthermore, it's crucial to acknowledge that the imbalanced distribution of entity types in the dataset will inevitably impact future results [8]. Labels with fewer examples will inherently be more difficult for the models to learn and predict accurately. One of the potential solutions for class imbalance is oversampling or undersampling. However, their application in the medical domain presents challenges due to the need for expert oversight to prevent errors during data augmentation. Hence, we'll continue working with the original dataset while acknowledging these limitations.

The imbalanced distribution of entity types in the frequency graph suggests that using accuracy as a metric may not be advisable. Accuracy could be skewed by the dominance of certain classes, potentially leading to misleading performance assessments. Accuracy is not recommended in most cases for NER tasks.

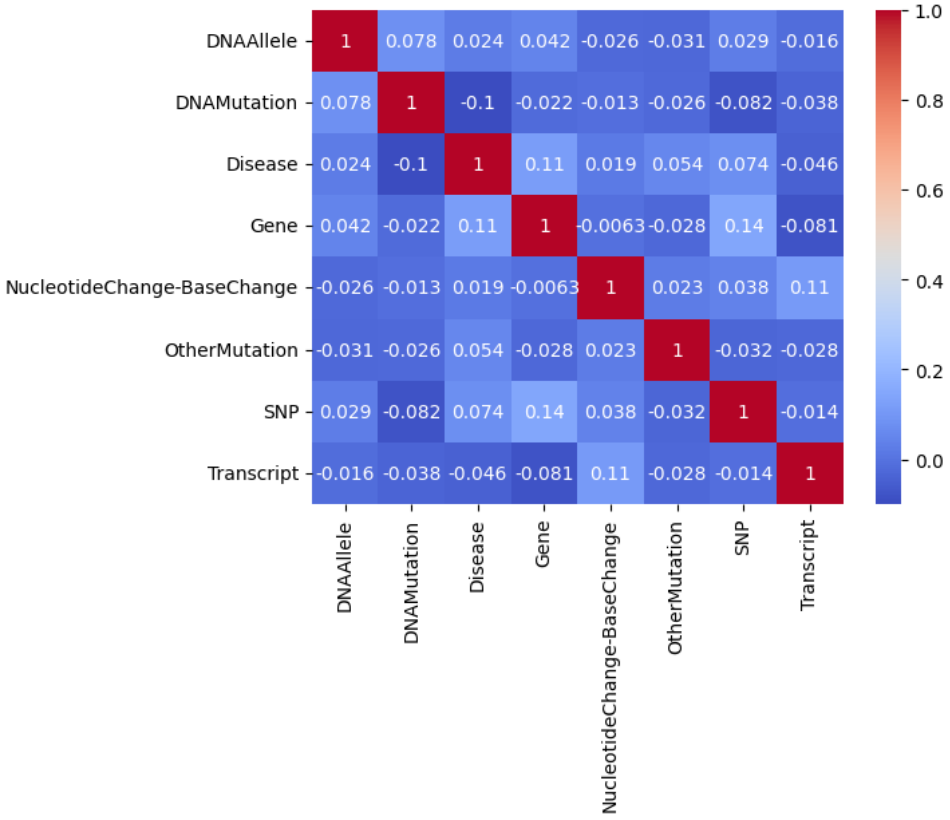

Figure 2: Correlation matrix of entity types.

The presence of predominantly low correlation values [Figure 2] suggests that the occurrence of certain entity types in the dataset is largely independent of each other. This implies that the identification or presence of one entity type does not necessarily influence the presence of another. However, positive correlations, such as those observed between "Disease" and "Gene" entities, indicate a potential association or co-occurrence between them. This aligns with their prevalence in the dataset. Conversely, negative correlations suggest a lack of association or potential mutual exclusivity between certain entity types, which could pose challenges in accurately identifying these entities in text. Additionally, such negative correlations can guide feature selection strategies, focusing on distinct linguistic patterns or contextual cues unique to each entity type.

3.2. Pre-processing

When it comes to preprocessing, it's essential to tailor our approach to the specific requirements and characteristics of the models we're working with. For instance, with GPT-3.5 Turbo, a model designed primarily for natural language understanding tasks, preprocessing is relatively straightforward due to its nature as a transformer-based language model. Since GPT is typically used for tasks like text generation or completion, there's no training involved on a specific dataset. Instead, we evaluate its performance directly on a test set.

Regarding RoBERTa, preprocessing involves converting single-tag entity labels into the BIO format commonly used for sequence labeling tasks. This requires tokenizing each clinical case, identifying entity boundaries, and assigning corresponding BIO labels to each token based on its position within the entity. Additionally, token indexing ensures alignment between tokens and their labels during model training or evaluation.

Conversely, models like GLiNER, which are explicitly designed for Named Entity Recognition (NER) tasks, offer more flexibility in preprocessing. GLiNER operates based on a token-level tagging approach, where each token in the input text is associated with a label indicating its entity type and its position within the entity (start and end offsets). Therefore, the preprocessing steps for GLiNER involve tokenizing the text and structuring the data to include the necessary information for each entity, such as its type and offsets.

3.3. Training details

It's worth noting that efforts have been made to allocate hyperparameters as fairly as possible for RoBERTa and GLiNER. However, exact parity couldn't be achieved due to variations in memory consumption between the models, a constraint imposed by the use of Google Colab. Despite these discrepancies, every endeavor has been made to ensure a balanced and equitable training setup for both models. Hyperparameters are listed in Table 1.

Table 1
Training hyperparameters used for our models.

Hyperparameter	RoBERTa	GLiNER
'num_train_epochs'	20	20
'per_device_train_batch_size'	16	2
'weight_decay'	0.01	0.01
'learning_rate'	2^{-5}	1^{-5}
'max_len'	700	700
'shuffle'	True	True
'freeze'	False	False

Due to differences in model architecture, tensor size, and computational operations, the batch size for GLiNER is substantially smaller compared to RoBERTa. While RoBERTa can accommodate batch sizes as large as 16 without issue, GLiNER's maximum usable batch size is restricted to 2, a reduction by a factor of four. This limitation necessitates adjustments to the learning rate. With smaller batch sizes, gradient estimation relies on fewer data points, potentially resulting in noisier estimations and slower convergence. Consequently, to mitigate these effects and stabilize training, a lower learning rate is required. Conversely, larger batch sizes typically facilitate faster convergence, warranting higher learning rates to ensure effective model updates [9].

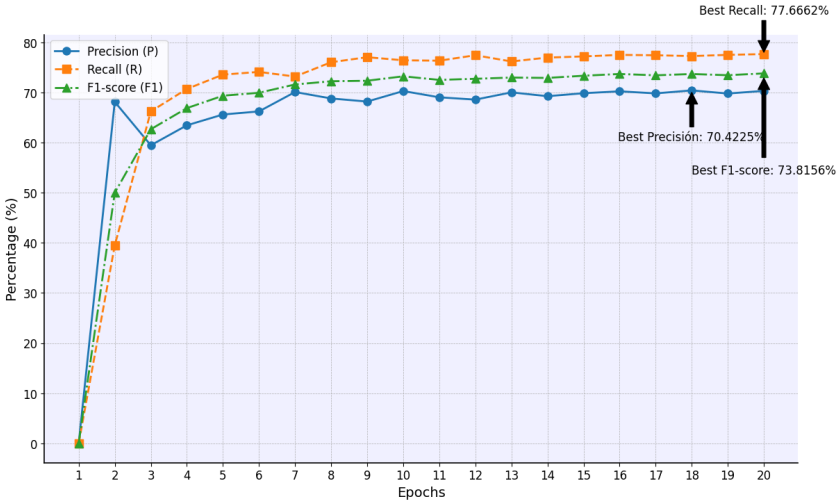

Figure 3: RoBERTa metrics evolution across epochs.

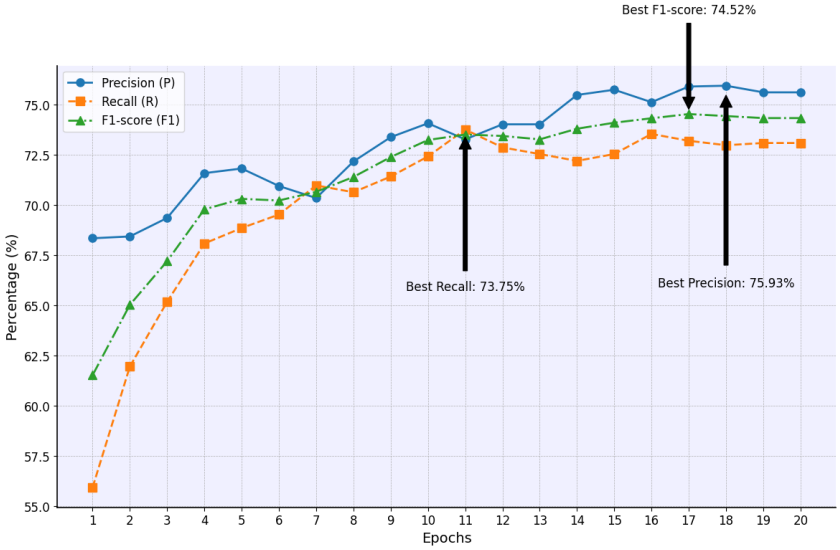

Figure 4: GLiNER metrics evolution across epochs.

Across the epochs, both models exhibit a consistent trend of improvement in precision, recall, and F1-score, suggesting effective learning and adaptation to the training data. However, when scrutinizing the specific metrics, several nuances emerge. RoBERTa demonstrates commendable performance, steadily increasing precision, recall, and F1-score over the epochs. Despite starting with lower values, it manages to narrow the performance gap with GLiNER as training progresses. This suggests that RoBERTa is capable of learning and refining its NER capabilities over time, albeit at a slower pace compared to GLiNER. On the other hand, GLiNER showcases superior performance across all metrics throughout the training process. It starts with higher precision, recall, and F1-score values compared to RoBERTa and maintains this lead consistently. The steady increase in performance metrics reflects the model's robustness and effectiveness in capturing and classifying named entities in biomedical text.

In conclusion, while both RoBERTa and GLiNER show promise for the NER task, GLiNER exhibits slightly superior performance in the context of our experiment.

Unlike RoBERTa and GLiNER, GPT models cannot undergo traditional fine-tuning processes due to their generative nature. Therefore, the approach relies on crafting effective prompts to guide the model towards accurate entity recognition. This unique challenge underscores the importance of

innovative prompt engineering strategies when utilizing GPT models for NER tasks. The initial attempts involved specifying the task, context, expected output format, and entity labels explicitly. However, these early prompts resulted in ambiguous outputs, often including additional labels or failing to adhere to the desired output format. Recognizing the need for a more nuanced approach, a series of iterative adjustments were made to the prompt, incorporating examples and refining the instructions to address specific issues encountered during model inference. Through meticulous experimentation, a final prompt was crafted that struck a balance between specificity and clarity. This refined prompt explicitly outlined the task, specified the expected entities, emphasized adherence to the desired output format, and provided instructions for handling cases with no detected entities. Despite the iterative nature of the process and the challenges posed by the generative nature of GPT models, the final prompt achieved satisfactory results, enabling the model to perform entity recognition with reasonable accuracy.

4. Results

The results of the Named Entity Recognition (NER) task evaluation are presented in this section [Table 2], showcasing the performance of the models.

Model	Precision	Recall	F1-Score
GPT-3.5 Turbo	0.400000	0.186578	0.254463
RoBERTa	0.611556	0.654926	0.632498
GLiNER Medium	0.790643	0.796287	0.793455

Table 2
Performance comparison of GPT-3.5, RoBERTa, and GLiNER in terms of Precision, Recall, and F1-Score.

The final results demonstrate clear distinctions in performance among the models evaluated. GPT-3.5 exhibits notably inferior performance compared to RoBERTa and GLiNER. This outcome aligns with expectations, considering GPT-3.5's autoregressive nature and lack of fine-tuning specifically for NER tasks, particularly within the intricate domain of medical text analysis. RoBERTa, pre-trained on medical texts, shows substantial improvement over GPT-3.5 across all metrics, indicating its ability to adapt effectively through fine-tuning. However, GLiNER Medium surpasses both models, delivering superior performance across all metrics. Its specialized architecture, designed and trained explicitly for NER tasks, enables more effective capture of relevant entity characteristics. Notably, it also demonstrates superior learning capability compared to RoBERTa, as evidenced by the availability of metric results [Figure 3 and Figure 4] from the initial epoch, suggesting a potentially faster convergence rate.

In terms of runtime efficiency, GPT-3.5 emerges as the fastest model, benefiting from its prompt-based inference approach without requiring pre-training. RoBERTa follows, with a per-epoch runtime of 45 seconds, while GLiNER exhibits the longest runtime at 2 minutes per epoch. These differences reflect the internal design, optimization, dataset size, and hyperparameters of each model [10]. Notably, GLiNER demonstrates superior learning capability compared to RoBERTa, as evidenced by the availability of metric results from the initial epoch, suggesting a potentially faster convergence rate.

To further justify the promising performance of GLiNER, we refer to a study conducted by the GLiNER team, as outlined in their paper [5]. In this study, the authors evaluated the zero-shot performance of various NER models [Figure 5], including GLiNER Large, ChatGPT, and UniNER-7B, across a diverse set of 20 datasets. Of particular relevance are datasets such as AnatEM, bc2gm, bc2chemd, bc5cdr, GENIA, and ncbi, all comprising biomedical literature data. Among these datasets, bc2gm, GENIA, and ncbi are most akin to our dataset, focusing on gene and protein annotations, molecular biology terms, and disease annotations, respectively. Notably, GLiNER achieved the highest F1-Score results for these datasets, with values of 47.9%, 55.5%, and 61.9%, respectively.

Dataset	ChatGPT	UniNER-7B	GLiNER-L
ACE05	26.6	**36.9**	27.3
AnatEM	30.7	25.1	**33.3**
bc2gm	40.2	46.2	**47.9**
bc4chemd	35.5	**47.9**	43.1
bc5cdr	52.4	**68.0**	66.4
Broad Tweeter	61.8	**67.9**	61.2
CoNLL03	52.5	**72.2**	64.6
FabNER	15.3	**24.8**	23.6
FindVehicle	10.5	22.2	**41.9**
GENIA	41.6	54.1	**55.5**
HarveyNER	11.6	18.2	**22.7**
MIT Movie	5.3	42.4	**57.2**
MIT Restaurant	32.8	31.7	**42.9**
MultiNERD	58.1	59.3	**59.7**
ncbi	42.1	60.4	**61.9**
OntoNotes	29.7	27.8	**32.2**
PolyglotNER	33.6	41.8	**42.9**
TweetNER7	40.1	**42.7**	41.4
WikiANN	52.0	55.4	**58.9**
WikiNeural	57.7	69.2	**71.8**
Average	36.5	45.7	**47.8**

Figure 5: Performance comparison of GLiNER, ChatGPT, and UniNER across a set of 20 NER datasets [5].

It is important to recognize that these datasets entail greater complexity compared to ours, resulting in lower performance metrics. This observation underscores the inherent challenge of entity detection, particularly within specific and complex domains such as biomedical text analysis. Thus, achieving high performance metrics, as demonstrated in our case, is not always feasible and highlights the difficulty associated with this task.

5. Conclusions

The culmination of this research effort underscores the critical role of Natural Language Processing (NLP) and Named Entity Recognition (NER) in current advancements. With a specific focus on the biomedical domain, the study has navigated through its complexities, leveraging recent datasets characterized by detailed entity labels. Evaluation of three models, including the established RoBERTa alongside the newer GPT-3.5 and GLiNER, revealed GLiNER's outstanding performance, surpassing its counterparts by over 15% in key metrics. This outcome underscores GLiNER's competitive edge and its promising trajectory in clinical applications. Furthermore, a comparative analysis shed light on the limitations of generative models when applied beyond their trained domains, offering valuable insights into their capabilities and constraints. In essence, this paper has significantly advanced our understanding of NER models' practical deployment, particularly in the medical domain, paving the way for enhanced entity identification and extraction, crucial for precise diagnostics and tailored treatments. According to official results, we secured second place in the competition, narrowly trailing behind the first-place finisher by less than a 3% margin across all three metrics. This achievement demonstrates the competitive performance of our proposed model.

References

[1] B. Guo, H. Liu, L. Niu, Integration of natural and deep artificial cognitive models in medical images: Bert-based ner and relation extraction for electronic medical records, Frontiers in Neuroscience 17 (2023) 1266771.

[2] U. Ahmed, K. Iqbal, M. Aoun, Natural language processing for clinical decision support systems:

A review of recent advances in healthcare, Journal of Intelligent Connectivity and Emerging Technologies 8 (2023) 1–16.

[3] M. M. Agüero-Torales, C. Rodríguez Abellán, M. Carcajona Mata, J. I. Díaz Hernández, M. Solís López, A. Miranda-Escalada, S. López-Alvárez, J. Mira Prats, C. Castaño Moraga, D. Vilares, L. Chiruzzo, Overview of GenoVarDis at IberLEF 2024: NER of Genomic Variants and Related Diseases in Spanish, Procesamiento del Lenguaje Natural 73 (2024).

[4] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[5] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, Gliner: Generalist model for named entity recognition using bidirectional transformer, 2023. arXiv:2311.08526.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[7] C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, M. Villegas, Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario, 2021. arXiv:2109.03570.

[8] Y. Mao, Y. Hao, W. Liu, X. Lin, X. Cao, Class-imbalanced-aware distantly supervised named entity recognition, IEEE Transactions on Neural Networks and Learning Systems (2023).

[9] S. L. Smith, P.-J. Kindermans, C. Ying, Q. V. Le, Don't decay the learning rate, increase the batch size, arXiv preprint arXiv:1711.00489 (2017).

[10] M. Li, T. Zhang, Y. Chen, A. J. Smola, Efficient mini-batch training for stochastic optimization, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 661–670.