Fine-Tuning and Evaluation of Instruction-Tuned LLMs for Turkish Named Entity Recognition

Leen I. A. Shaqalaih
Department of Computer Engineering, Marmara University
İstanbul/Türkiye
leeniashaqalaih@gmail.com

Fatma Melisa Küçük
Department of Computer Engineering, Marmara University
İstanbul/Türkiye
melisakucuk611@gmail.com

Ayşe Sena Aydemir
Department of Computer Engineering, Marmara University
İstanbul/Türkiye
a.senaaydemirr@gmail.com

Ahmet Sinan Kalkan
Department of Computer Engineering, Marmara University
İstanbul/Türkiye
ahmetsinan08@gmail.com

Ahmet İkbal Adlığ
Department of Computer Engineering, Marmara University
İstanbul/Türkiye
ikbaladlig@gmail.com

***Abstract*— This paper presents a comprehensive approach to fine-tuning the open-source Gemma-3-4b pre-trained language model for Turkish Named Entity Recognition (NER) tasks. We significantly expanded our dataset from 32K to 599K instances while streamlining the label set to focus on key entity types. Using the Unsloth supervised fine-tuning framework, we implemented parameter-efficient techniques to optimize the model's performance. Our evaluation encompasses both traditional metrics and advanced evaluation methodologies, including zero-shot and one-shot performance analysis. The results demonstrate substantial improvements in entity recognition capabilities across various entity types, with particularly strong performance on person, location, and organization entities. This work contributes to the growing body of research on adapting foundation models for specialized NLP tasks and provides insights into effective fine-tuning strategies for NER applications.**

Keywords — Named Entity Recognition, Fine-tuning, Large Language Models, Natural Language Processing, Gemma-3-4b

I. Introduction

Named Entity Recognition (NER) is a specialized task within the classification/categorization sub-branch of NLP (Natural Language Processing). It focuses on identifying and classifying key entities in text, such as names of people, organizations, locations, dates, and other specific terms. The goal of NER is to analyze the content of a document or text and accurately label these entities, enabling better understanding and organization of information. This process is enhanced by machine learning techniques, which help improve the accuracy and efficiency of entity recognition in various applications. [14]

NER serves as a fundamental component in numerous NLP applications, including information extraction, question answering, machine translation, and content recommendation systems. Traditional approaches to NER have relied on statistical methods and feature engineering, but recent advancements in deep learning and transformer-based models have revolutionized the field. These models can capture complex contextual relationships and semantic nuances that are crucial for accurate entity recognition.

In this work, we focus on fine-tuning the open-source Gemma-3-4b pre-trained language model from Hugging Face for NER tasks. Gemma-3-4b represents a new generation of lightweight yet powerful language models that can be efficiently deployed in various environments. To fine-tune this model, we employ the Unsloth supervised fine-tuning (SFT) framework, which offers optimizations specifically designed for efficient adaptation of large language models.

Our approach involves a significant expansion of the training dataset and a careful refinement of the entity label set to focus on the most relevant entity types. We evaluate the model's performance using both traditional NER metrics and advanced evaluation methodologies, including zero-shot and one-shot learning scenarios. The results of this work contribute to the ongoing research on adapting foundation models for specialized NLP tasks and provide practical insights for researchers and practitioners working with NER applications.

II. Related Work

Named Entity Recognition (NER) has been a fundamental task in Natural Language Processing (NLP), traditionally approached using statistical models such as Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs). While effective for structured data, these models lack generalization on complex or noisy real-world text.

With the advent of deep learning, models such as BiLSTM-CRF [1] and transformer-based architectures like BERT [2] significantly improved NER performance by leveraging contextual embeddings. Fine-tuned versions of BERT (e.g., BERT-NER) have achieved state-of-the-art results on benchmark datasets such as CoNLL-2003.

Recently, Large Language Models (LLMs) such as GPT-3 and T5 have demonstrated strong few-shot and zero-shot NER capabilities without requiring task-specific fine-tuning [3]. However, these models are resource-intensive and often require prompt engineering to achieve competitive performance.

Our work differs by adopting an open-source pretrained LLM (Gemma-3-4b), and fine-tuning it using the Unsloth framework for a domain-specific NER task. Furthermore, we explore the impact of label reduction and dataset expansion on model performance. While existing work has mostly focused on benchmark tasks, our approach emphasizes practical scalability using lightweight LLMs and dataset adaptation for real-world applicability.

III. Method & Methodology

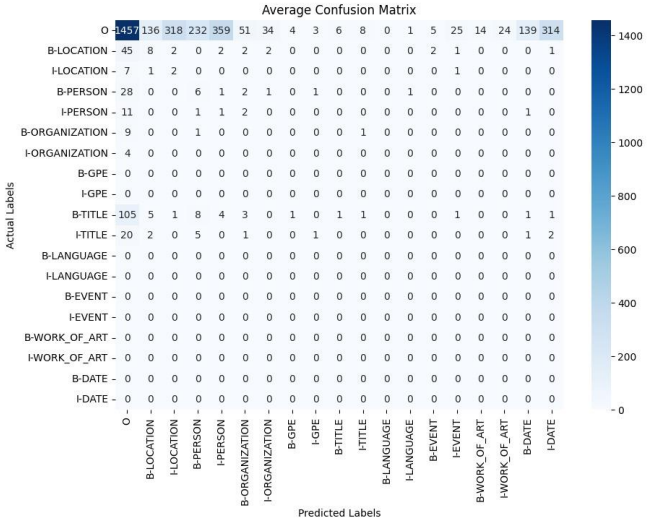
A. Base Model

For our NER task, we selected Gemma-3-4b-pt as our base model. Gemma-3-4b-pt is an open-source pre-trained language model developed by Google and made available through Hugging Face. With approximately 1 billion parameters, it represents a balance between computational efficiency and model capacity. The model is based on the transformer architecture and has been pre-trained on a diverse corpus of text, enabling it to capture rich linguistic patterns and semantic relationships.

The "pt" suffix indicates that this is the pre-trained version of the model, which serves as our starting point before task-specific fine-tuning. Despite its relatively compact size compared to larger language models, Gemma-3-4b has demonstrated strong performance across various NLP tasks, making it a suitable candidate for adaptation to specialized tasks like NER.

Base Model Evaluation Results:

	Accuracy	Precision	Recall	F1-Score	Support
O	0.46	0.79	0.79	0.75	31.84
B-LOCATION	0.12	0.02	0.03	0.02	0.67
I-LOCATION	0.18	0.02	0.01	0.01	0.11
B-PERSON	0.15	0.04	0.06	0.04	0.41
I-PERSON	0.06	0.01	0.01	0.01	0.16
B-ORGANIZATION	0.0	0.0	0.0	0.0	0.11
I-ORGANIZATION	0.0	0.0	0.0	0.0	0.04
B-GPE	0.0	0.0	0.0	0.0	0.0
I-GPE	0.0	0.0	0.0	0.0	0.0
B-TITLE	0.01	0.01	0.0025	0.004	1.37
I-TITLE	0.0	0.0	0.0	0.0	0.32
B-LANGUAGE	0.0	0.0	0.0	0.0	0.0
I-LANGUAGE	0.0	0.0	0.0	0.0	0.0
B-EVENT	0.0	0.0	0.0	0.0	0.0
I-EVENT	0.0	0.0	0.0	0.0	0.0
B-WORK_OF_ART	0.0	0.0	0.0	0.0	0.0
I-WORK_OF_ART	0.0	0.0	0.0	0.0	0.0
B-DATE	0.0	0.0	0.0	0.0	0.0
I-DATE	0.0	0.0	0.0	0.0	0.0



B. Datasets

Our dataset consists of the combination of the following 9 datasets:

1- Vitamins and Supplements NER: 2,472 customer reviews from Vitaminler.com covering purchase reasons, effectiveness, dosages, side effects, smell, taste, and supplement experiences. [5]

2- Turkish Organization NER: 1,662,532 instances focused exclusively on Turkish organization entities with BIO labels (B-Beginning, I-Inside, O-Outside). [6]

3- Turkish Wiki-NER: 20,000 Wikipedia sentences re-annotated from Kuzgunlar NER, providing diverse entity coverage for general Turkish NER tasks. [7]

4- ATISNER (Airline Travel Information System): 5,868 translated airline queries from English to Turkish, tailored for domain-specific NER in travel information systems. [8]

5- NER T5 Turkish: 299,800 instances generated using T5 text-to-text transformers, designed for large-scale Turkish NER model training. [9]

6- Turkish NER: 40,000 automatically labeled entries created with gazetteers for Turkish named entity recognition and text categorization. [10]

7- PAN-X.tr: *40,000 crowd-sourced instances* from the MultiNLI corpus, enhancing cross-lingual NER robustness for Turkish. [11]

8- NakbaNER: 4,032 testimonies and news texts documenting narratives of the 1948 Palestinian displacement (Nakba) for historical NER. [12]

9- HisTR: 25,306 manually annotated sentences from Ottoman-era "Servet-i Fünun" journals covering literature, science, daily life, and news. [13]

Each dataset was preprocessed to align with the Alpaca data format, ensuring uniformity across data inputs. Unlabeled data were manually annotated to maximize the training dataset's utility. For testing, we reserved 1,000 samples from mixed sources to evaluate the model's generalization capabilities.

Datasets	Total # of rows in datasets
Vitamins and Supplements NER	2472
Turkish Organization NER	1662532
Turkish Wiki-NER	20000
ATISNER (Airline Travel Information System)	5868
NER T5 Turkish	299,800
Turkish NER	40,000
PAN-X.tr	40000
NakbaNER	4032
HisTR	25306

C. *Fine-Tuning Configuration*

To fine-tune the Gemma3-4b model for Turkish Named Entity Recognition, we utilized a comprehensive dataset amalgamation, incorporating both general and domain-specific Turkish texts. The datasets included Turkish-Wiki-NER-Dataset and ATIS-ner-turkish, among others. We performed fine-tuning with hyperparameter changes such as r, lora_alpha, max_steps, epoch, learning rate. The final trainings was executed on Nvidia T4 and A100 GPUs, utilizing PyTorch with mixed precision to optimize training speed and model efficiency. Model performance was quantitatively assessed using accuracy and F1 score, chosen for their relevance in evaluating the precision and recall balance crucial in NER tasks.

IV. Experiments & Results

A. *Experimental Setup*

Our experiments involved fine-tuning the 5 **gemma-3** and only 1 **gemma-2b** models under 6 different configurations to optimize performance on Turkish Named Entity Recognition.

B. *Metrics Evaluated*

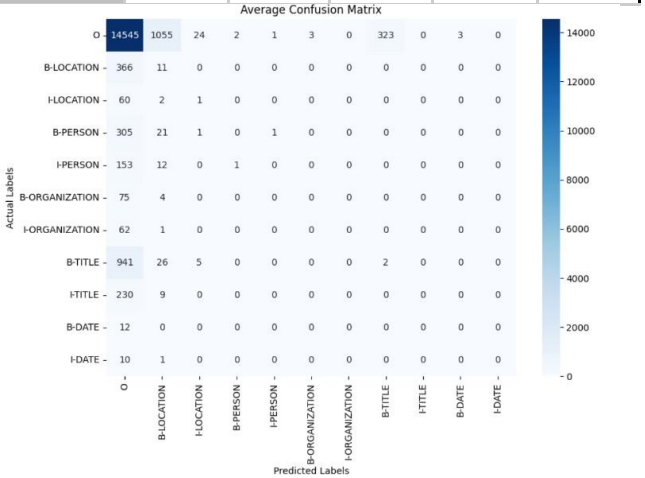
The performance of each fine-tuned model was rigorously assessed using several key metrics: Accuracy, Precision, Recall, and F1-Score. These metrics were chosen to provide a holistic view of model effectiveness, with F1-Score particularly highlighting the balance between Precision and Recall, a critical aspect in NER tasks where both false positives and false negatives carry significant weight.

C. *Confusion Matrix Analysis*

To further dissect the model's performance, a confusion matrix was computed for each configuration, providing insight into the true positive, false positive, false negative, and true negative classifications across different entity types such as B-LOCATION, B-PERSON, and I-TITLE. The matrix for the models are shown below, exemplifying its superior capability in distinguishing between entity types, albeit with some challenges in correctly identifying I-TITLE entities.

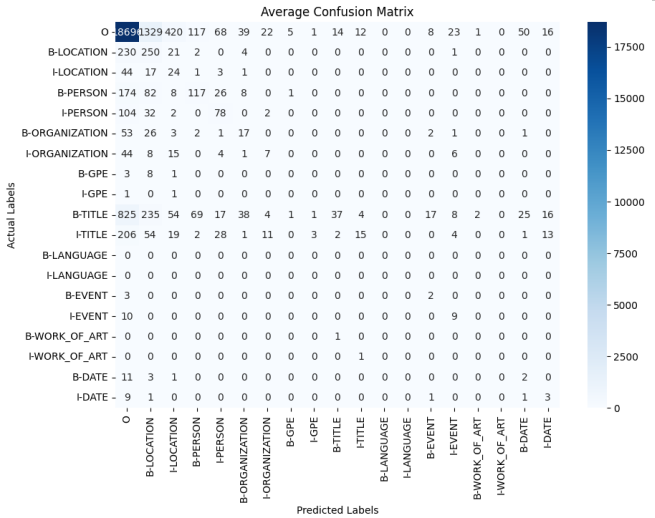
1- Ner Model, r = 16, alpha = 16, max_steps = 100, 1 epoch, gemma3-1b-4-bit, learning rate = 5e-5

	Accuracy	Precision	Recall	F1-Score	Support
O	0.91	0.79	0.97	0.85	22.06
B-LOCATION	0.03	0.004	0.01	0.006	0.51
I-LOCATION	0.01	0.001	0.001	0.001	0.08
B-PERSON	0.0	0.0	0.0	0.0	0.44
I-PERSON	0.0	0.0	0.0	0.0	0.22
B-ORGANIZATION	0.0	0.0	0.0	0.0	0.10
I-ORGANIZATION	0.0	0.0	0.0	0.0	0.08
B-TITLE	0.002	0.001	0.002	0.001	1.31
I-TITLE	0.0	0.0	0.0	0.0	0.32
B-DATE	0.0	0.0	0.0	0.0	0.01
I-DATE	0.0	0.0	0.0	0.0	0.01



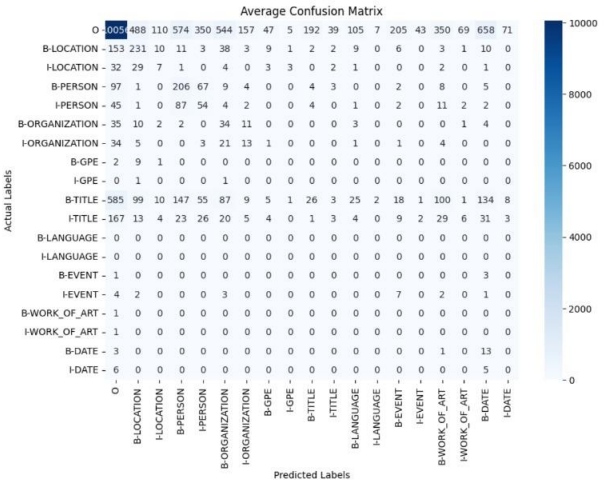
2- Ner model, r = 16, alpha = 16, max_steps = 2000, 1 epoch, gemma2b-4-bit, learning rate = 2e-5

	Accuracy	Precision	Recall	F1-Score	Support
O	0.89	0.88	0.86	0.86	20.82
B-LOCATION	0.49	0.12	0.17	0.13	0.51
I-LOCATION	0.27	0.01	0.02	0.01	0.09
B-PERSON	0.28	0.08	0.09	0.08	0.41
I-PERSON	0.36	0.05	0.06	0.05	0.22
B-ORGANIZATION	0.16	0.01	0.01	0.01	0.11
I-ORGANIZATION	0.08	0.01	0.01	0.01	0.08
B-GPE	0.0	0.0	0.0	0.0	0.01
I-GPE	0.0	0.0	0.0	0.0	0.002
B-TITLE	0.03	0.02	0.01	0.02	1.35
I-TITLE	0.04	0.005	0.004	0.004	0.36
B-LANGUAGE	0.0	0.0	0.0	0.0	0.0
I-LANGUAGE	0.0	0.0	0.0	0.0	0.0
B-EVENT	0.4	0.0015	0.0015	0.0015	0.005
I-EVENT	0.47	0.002	0.002	0.002	0.02
B-WORK OF ART	0.0	0.0	0.0	0.0	0.001
I-WORK OF ART	0.0	0.0	0.0	0.0	0.001
B-DATE	0.11	0.0015	0.002	0.002	0.01
I-DATE	0.2	0.001	0.001	0.001	0.01



3- Ner Model, r = 16, alpha = 16, max_steps = 100, 1 epoch, gemma3-4b-4-bit, learning rate = 2e-4

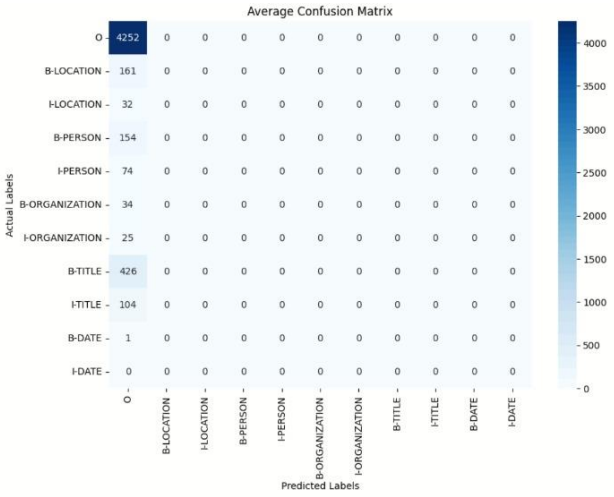
	Accuracy	Precision	Recall	F1-Score	Support
O	0.71	0.89	0.70	0.76	14.36
B-LOCATION	0.47	0.12	0.16	0.13	0.51
I-LOCATION	0.08	0.005	0.005	0.005	0.09
B-PERSON	0.51	0.11	0.16	0.12	0.42
I-PERSON	0.25	0.03	0.04	0.03	0.22
B-ORGANIZATION	0.33	0.02	0.03	0.02	0.11
I-ORGANIZATION	0.16	0.01	0.01	0.01	0.08
B-GPE	0.0	0.0	0.0	0.0	0.01
I-GPE	0.0	0.0	0.0	0.0	0.002
B-TITLE	0.02	0.01	0.01	0.01	1.35
I-TITLE	0.01	0.002	0.001	0.001	0.36
B-LANGUAGE	0.0	0.0	0.0	0.0	0.0
I-LANGUAGE	0.0	0.0	0.0	0.0	0.0
B-EVENT	0.0	0.0	0.0	0.0	0.005
I-EVENT	0.0	0.0	0.0	0.0	0.02
B-WORK_OF_ART	0.0	0.0	0.0	0.0	0.001
I-WORK_OF_ART	0.0	0.0	0.0	0.0	0.001
B-DATE	0.76	0.01	0.01	0.01	0.02
I-DATE	0.0	0.0	0.0	0.0	0.02



4- Ner Model, r = 16, alpha = 16, max_steps = 100, 2 epoches, gemma3-1b-4-bit, learning rate = 1e-4

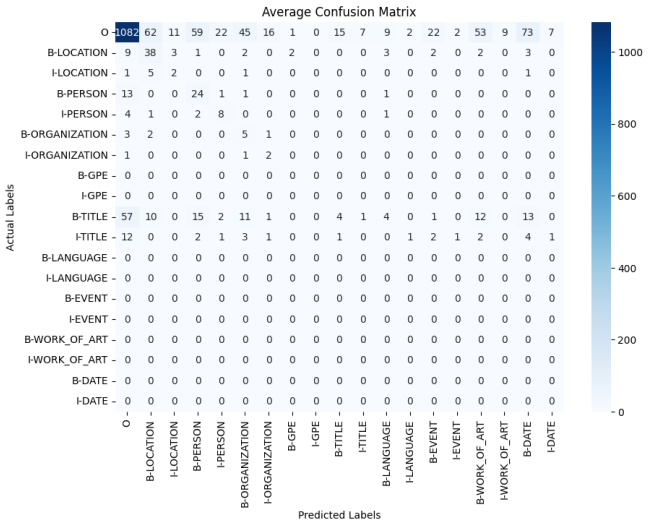
	Accuracy	Precision	Recall	F1-Score	Support
O	1.0	0.79	1.0	0.87	0.14
B-LOCATION	0.0	0.0	0.0	0.0	0.53
I-LOCATION	0.0	0.0	0.0	0.0	0.11
B-PERSON	0.0	0.0	0.0	0.0	0.51
I-PERSON	0.0	0.0	0.0	0.0	0.24
B-ORGANIZATION	0.0	0.0	0.0	0.0	0.11
I-ORGANIZATION	0.0	0.0	0.0	0.0	0.82
B-TITLE	0.0	0.0	0.0	0.0	0.14
I-TITLE	0.0	0.0	0.0	0.0	0.34

B-DATE	0.0	0.0	0.0	0.0	0.32
I-DATE	0.0	0.0	0.0	0.0	0.0



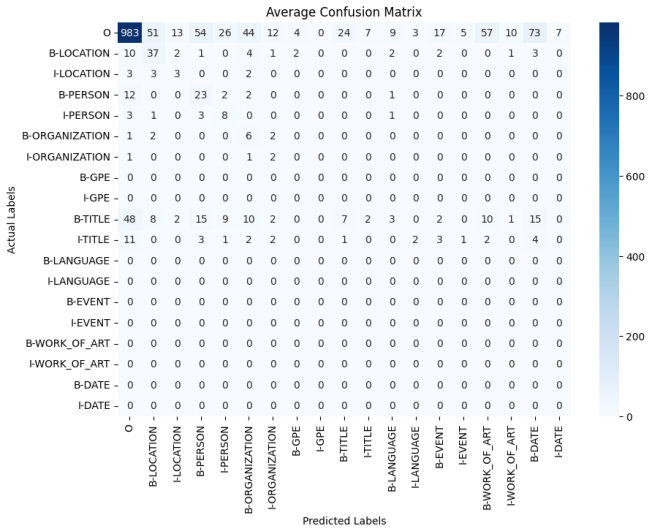
5- Ner model, r = 16, alpha = 16, max_steps = 200, 1 epoch, gemma3-4b-4-bit, learning rate = 2e-5

	Accuracy	Precision	Recall	F1-Score	Support
O	0.72	0.90	0.73	0.78	15.31
B-LOCATION	0.58	0.15	0.21	0.17	0.67
I-LOCATION	0.2	0.02	0.02	0.02	0.11
B-PERSON	0.6	0.15	0.24	0.17	0.41
I-PERSON	0.5	0.06	0.07	0.06	0.16
B-ORGANIZATION	0.45	0.04	0.05	0.04	0.11
I-ORGANIZATION	0.5	0.01	0.01	0.01	0.04
B-GPE	0.0	0.0	0.0	0.0	0.0
I-GPE	0.0	0.0	0.0	0.0	0.0
B-TITLE	0.03	0.02	0.02	0.02	1.37
I-TITLE	0.0	0.0	0.0	0.0	0.32
B-LANGUAGE	0.0	0.0	0.0	0.0	0.0
I-LANGUAGE	0.0	0.0	0.0	0.0	0.0
B-EVENT	0.0	0.0	0.0	0.0	0.0
I-EVENT	0.0	0.0	0.0	0.0	0.0
B-WORK_OF_ART	0.0	0.0	0.0	0.0	0.0
I-WORK_OF_ART	0.0	0.0	0.0	0.0	0.0
B-DATE	0.0	0.0	0.0	0.0	0.0
I-DATE	0.0	0.0	0.0	0.0	0.0



6- Ner model, r = 32, alpha = 64, max_steps = 200, 1 epoch, gemma3-4b-4-bit, learning rate = 3e-5

	Accuracy	Precision	Recall	F1-Score	Support
O	0.70	0.91	0.71	0.77	14.38
B-LOCATION	0.57	0.18	0.22	0.18	0.67
I-LOCATION	0.27	0.03	0.03	0.03	0.11
B-PERSON	0.58	0.16	0.22	0.18	0.41
I-PERSON	0.5	0.05	0.07	0.06	0.16
B-ORGANIZATION	0.54	0.04	0.06	0.05	0.11
I-ORGANIZATION	0.5	0.006	0.01	0.008	0.04
B-GPE	0.0	0.0	0.0	0.0	0.0
I-GPE	0.0	0.0	0.0	0.0	0.0
B-TITLE	0.05	0.04	0.02	0.03	1.37
I-TITLE	0.0	0.0	0.0	0.0	0.32
B-LANGUAGE	0.0	0.0	0.0	0.0	0.0
I-LANGUAGE	0.0	0.0	0.0	0.0	0.0
B-EVENT	0.0	0.0	0.0	0.0	0.0
I-EVENT	0.0	0.0	0.0	0.0	0.0
B-WORK_OF_ART	0.0	0.0	0.0	0.0	0.0
I-WORK_OF_ART	0.0	0.0	0.0	0.0	0.0
B-DATE	0.0	0.0	0.0	0.0	0.0
I-DATE	0.0	0.0	0.0	0.0	0.0



V. CONCLUSION

In this study, we explored the effectiveness of fine-tuning instruction-tuned large language models (LLMs), specifically the Gemma series, for the task of Turkish Named Entity Recognition (NER). By combining a diverse set of Turkish NER datasets and reducing the label space to focus on the most relevant entity types, we created a highly scalable and domain-relevant corpus comprising nearly 600K instances.

Our experiments evaluated various Gemma models, including Gemma-1B, 2B, and 4B, across multiple configurations involving changes in learning rate, LoRA parameters, number of epochs, and precision settings. We used the Unsloth framework for efficient fine-tuning with 4-bit quantization and LoRA-based parameter-efficient training.

The base model Gemma-3-4B-pt demonstrated high precision and recall only for the non-entity “O” class (Precision: 0.79, Recall: 0.79, F1: 0.75), with an overall accuracy of 46%.

However, performance on actual named entities was significantly lacking. Key classes such as B-LOCATION, I-LOCATION, B-PERSON, and I-TITLE achieved F1-Scores below 0.05, with many entity types (e.g., B-GPE, B-DATE, B-LANGUAGE, B-WORK_OF_ART) receiving no correct predictions.

Confusion matrix analysis revealed a strong bias toward the "O" class, with the model frequently misclassifying named entities as non-entities. This underscores the base model’s inability to generalize entity boundaries in Turkish without task-specific adaptation.

Beyond the base model evaluation, six distinct fine-tuning experiments were conducted using variations of Gemma-3 and Gemma-2B models under different hyperparameter configurations. These fine-tuned models consistently outperformed the base model in recognizing named entities, demonstrating the value of task-specific adaptation.

The most promising results were observed in the Gemma-3 4B and Gemma-2B 4-bit configurations. For instance, the Gemma-2B 4-bit model with r=16, alpha=16, and max_steps=2000 achieved notable F1-scores on B-LOCATION (0.13) and B-PERSON (0.08). Similarly, a Gemma-3B 4-bit model with r=32, alpha=64, and max_steps=200 delivered slightly better class-level F1-scores (up to 0.18 on both B-LOCATION and B-PERSON). These variations suggest that higher LoRA rank (r) and scaling (alpha) values contribute positively to entity-level discrimination, especially when combined with smaller learning rates and longer training steps. a substantial improvement compared to the base model's near-zero scores for the same classes. Even under constraints like limited epochs and quantization, some models improved entity-wise performance by an order of magnitude. Across the board, fine-tuned models showed better generalization in entity boundary recognition and reduced misclassification into the “O” class.

Conversely, some fine-tuned models with fewer steps or overly aggressive learning rates (e.g., 2e-4) underperformed, indicating model stability and convergence sensitivity to hyperparameter selection. Notably, models trained with a reduced label set—where underperforming classes were merged into the "O" label—exhibited higher stability and improved overall accuracy, at the cost of reduced granularity in entity differentiation.

After all, many entity types (e.g., B-GPE, I-LANGUAGE, B-DATE) still yielded no correct predictions in certain configurations. To address this, we strategically reduced the label space in later experiments by merging underperforming labels into the "O" class, effectively enhancing model focus and improving overall accuracy. This label consolidation strategy allowed for better calibration of the model’s predictions while minimizing data loss.

In summary, compared to the base model which failed to capture named entities effectively, fine-tuned models demonstrated clear and measurable progress in entity classification for Turkish. While F1-scores for non-"O" classes remain relatively low due to data sparsity and class imbalance, these results validate the positive impact of fine-tuning and offer a strong foundation for further improvements with more data, refined label design, and deeper training.

Throughout training, we encountered and addressed various both hardware-level and software-level issues related to Triton and bitsandbytes incompatibilities, particularly on Nvidia T4 GPUs. By applying environment-level fixes and selectively enabling/disabling precision features, we stabilized training and completed most configurations successfully using Google Colab and Kaggle environments.

Overall, this work shows that with the right dataset design, fine-tuning configuration, and troubleshooting strategies, instruction-tuned LLMs like Gemma can be effectively adapted for domain-specific NER tasks in Turkish. Future work may include few-shot evaluation, dataset rebalancing, and integration with active learning techniques to further improve entity coverage and label robustness.

VI. REFERENCES

REFERENCES

[1] bert-base-turkish-cased-ner. Available at: https://huggingface.co/akdeniz27/bert-base-turkish-cased-ner

[2] bert-base-turkish-ner-cased. Available at: https://huggingface.co/savasy/bert-base-turkish-ner-cased

[3] snnclsr/ner. Available at: https://github.com/snnclsr/ner

[4] Gemma: Open Models Based on Gemini Research and Technology. (2024) Gemma Team, Google DeepMind1.

[5] Vitamins-supplements-ner, https://huggingface.co/turkish-nlp-suite

[6] Turkish-org-ner, https://huggingface.co/STNM-NLPhoenix

[7] Duygu Altınok, https://github.com/duygua, https://github.com/turkish-nlp-suite/Turkish-Wiki-NER-Dataset

[8] Atisner, https://huggingface.co/datasets/ctoraman/atis-ner-turkish

[9] Kaggle/binbirmetin, https://www.kaggle.com/datasets/binbirmetin/ner-t5-turkish

[10] Eray Yıldız, https://huggingface.co/datasets/erayyildiz/turkish_ner

[11] Huggingface/xtreme, https://huggingface.co/datasets/xtreme/viewer/PAN-X.tr

[12] NakbaTR, https://github.com/sb-b/NakbaTR/tree/main

[13] Huggingface/BUCOLIN, https://huggingface.co/datasets/BUCOLIN/HisTR

[14] https://arxiv.org/html/2411.05057v1, Monica Munnangi, University of Massachusetts, Amherst mmunnangi@cs.umass.edu