Evaluating Large Language Models for Medical Named Entity Recognition in Urdu: A Benchmark Study

Bushra Nasim[1], Kinza Latif[1], Muhammad Zohair[1], Muhammad Hassan Asif[1], Zarmeen Nasim[2]

[1] Dow Medical College, Karachi, Pakistan, [2] CITRIC Health Data Science Centre, Aga Khan University, Karachi, Pakistan
{bushra.nasim02, kinzalatif01, zohair.ajaz, muhammadhassanasif02}@gmail.com, zarmeen.nasim@aku.edu

Abstract

Medical named entity recognition (NER) is a crucial task in natural language processing (NLP) for extracting meaningful entities such as diseases, symptoms, medications, body parts, and treatments from clinical text. However, NER in low-resource languages like Urdu remains underexplored due to limited annotated datasets. In this study, we evaluated the performance of two state-of-the-art large language models (LLMs), ChatGPT-4o and LLAMA 3.2, on Urdu medical NER using a dataset of 2,057 health-related Urdu news headlines manually annotated across five entity categories. Both models were evaluated using precision, recall, and F1-score. It was found that both models exhibited low precision and moderate recall. ChatGPT-4o achieved the highest F1 for Disease (0.35) while LLAMA 3.2 reached slightly lower F1 scores for Disease (0.33). Both models performed poorly on treatment-related terms, with F1 scores of 0.036 (LLAMA 3.2) and 0.011 (ChatGPT-4o). Micro-average F1-scores were 0.187 for ChatGPT-4o and 0.183 for LLAMA 3.2, indicating comparable overall performance. These findings highlight the challenges of medical NER in low-resource languages and underscore the need for domain-specific fine-tuning, transfer learning, few-shot learning, and prompt engineering to improve performance.

Keywords: medical NER, low resource language, Urdu language processing, LLM, healthcare, medical entities

1. Introduction

In natural language processing (NLP), Named Entity Recognition (NER) is a fundamental task aimed at determining and categorizing entities within the given text. The task is crucial for various domains including healthcare. In the context of healthcare domain, efficient and accurate labelling of named entities enables a wide range of applications, including clinical record summarization, automated extraction of symptoms and diagnoses, pharmacovigilance, and public health surveillance. Recent advances in large language models (LLMs) have made significant improvements across various NLP tasks such as sentiment analysis(Deng et al., 2023), machine translation(Coleman et al., 2024; Elshin et al., 2024), NER(Iwakami et al., 2024), text generation(Kumichev et al., 2024) high-resource languages. Nonetheless, the applicability of these models to low-resource languages remains underexplored. Most LLMs are pre-trained predominantly on English and other high resources languages, resulting in performance disparities when applied to low resource languages.

Low-resource languages are often characterized by limited digital data, annotated resources and preprocessing tools. Urdu, a national language of Pakistan and spoken by 300 million people (Nasim & Haider, 2022) worldwide, is considered a low resource language with limited labeled data available specifically for medical domains. Medical NER in Urdu presents additional linguistic and technical challenges. These include rich morphology, free word order, absence of capitalization cues for entity boundaries and dialectal variation. Such characteristics complicate entity recognition and reduce the effectiveness of models trained primarily in high-resource languages.

Recent studies have explored the capability of large language models (LLMs) in English language for medical NER, reporting promising results in zero-shot or few-shot settings (Jahan et al., 2024). However, their effectiveness in specialized domains and low-resource settings like Urdu medical text remains unclear. Systematic benchmarking in this context is therefore necessary to understand both their potential and their limitations.

In this study, we aim to evaluate the performance of the two state-of-the-art LLMs: ChatGPT-4o and LLAMA 3.2 on medical NER task in Urdu language. We focus on identifying five critical entity categories: diseases, symptoms, medications, treatments, and body parts. These entities are central to healthcare text analytics and have direct implications for clinical decision support and public health monitoring. To facilitate this evaluation, we created a manually annotated dataset of 2,057 Urdu health-related news headlines. The primary objective of this work is to benchmark the zero-shot performance of general-purpose LLMs on Urdu medical NER and to analyze the challenges encountered in this low-resource, morphologically rich language setting.

The major contributions of this study are as follows:

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 40–48

- To the best of our knowledge, we present the first manually annotated dataset specifically designed for medical NER in Urdu.
- We provide systematic benchmarking of two state-of-the-art LLMs on Urdu medical NER, offering quantitative and qualitative insights into their strengths and limitations.
- We analyze the linguistic and domain-specific challenges that affect LLM performance in a low-resource South Asian language, highlighting implications for future research in healthcare NLP.

2. Related Work

Named entity recognition (NER) is an important domain of Natural Language Processing (NLP) with increasing applications in the medical domain. Various methodologies have been employed for medical NER, ranging from traditional rule-based systems to advanced deep learning approaches. While significant progress has been made in high-resource languages like English, the development of medical NER in low-resource languages such as Urdu, remains underexplored. This section reviews advancements in medical NER, focusing on the challenges in low-resource languages and the emerging role of LLMs in addressing these challenges.

In high-resource languages, medical NER has benefited significantly from advances in deep learning and pre-trained models. Models like BERT-based BioBERT(Lee et al., 2020a) and ClinicalBERT (Alsentzer et al., 2019), trained on biomedical corpora, have achieved state-of-the-art performance by leveraging domain-specific embeddings to capture complex medical terminology (Lee et al., 2020b). Other approaches, such as Conditional Random Field (CRF) (Zhao et al., 2012) and BiLSTM-CRF architectures (Nasim et al., 2020) , have been used for sequence labeling tasks and remain popular due to their ability to incorporate contextual dependencies (Lample et al., 2016). Despite these advances, direct application of such models to low-resource languages like Urdu is limited by the lack of annotated datasets and domain-specific pre-training.

LLMs such as GPT-3 (Zhang & Li, 2021), PaLM(Anil et al., 2023), and ChatGPT(Wu et al., 2023) have demonstrated remarkable performance in general NLP tasks, including NER (Brown et al., 2020). These models are trained on massive multilingual datasets, which allow them to generalize across languages and domains. Efforts to adapt LLMs for NER include techniques like LoRA (Low-Rank Adaptation), which fine-tunes pre-trained models efficiently for domain-specific tasks (Wang et al., 2023). While these methods show promise, integrating the generative outputs of LLMs into structured tasks such as NER remains an ongoing challenge, particularly in low-resource settings.

NER in low-resource languages such as Urdu faces unique challenges, such as lack of capitalization, free word order, morphological richness, a context-dependent lexicon, and the scarcity of training data(Naz et al., 2014). Most of the work on Urdu NER has focused on general domains rather than specialized fields like medicine. For example, Anam et al. (Anam et al., 2024) proposed using Floret embeddings combined with BiLSTM, GRU, and CRF for general Urdu NER.

However, these models often fall short in domain-specific tasks like medical NER due to the lack of specialized training data in the target language. Medical texts are inherently complex, containing domain-specific terminologies, abbreviations, lack of formatting, missing punctuations, misspelling that require a nuanced understanding of the language and its medical semantics(Leaman et al., 2015). The challenge is amplified in low-resource languages like Urdu, where limited annotated datasets and pretraining corpora are available for domain adaptation. Researchers have explored hybrid methods, such as combining rule-based approaches with machine learning techniques, but these approaches are less adaptable to the complex structure of medical text (Ahmed et al., 2024).

Research in NER for Urdu, a low-resource language, has largely focused on general domains. Recent studies have highlighted the limitations of these approaches in capturing domain-specific context, especially for medical NER, where abbreviations, synonyms, and code-switching are common. Furthermore, the lack of annotated datasets in Urdu for medical text is a critical challenge. High-quality labeled datasets are required for training and evaluating advanced deep learning models. Nonetheless, developing such datasets is labor-intensive and requires domain expertise. For Urdu, the scarcity of annotated corpora in the medical domain significantly hinders progress. Unlike high-resource languages such as English, where extensive annotated datasets exist for medical text, Urdu suffers from a lack of standardized resources.

To the best of our knowledge, this study is the first attempt of evaluating the performance of LLMs for medical NER in Urdu language, highlighting a significant gap in the existing literature. Additionally, our study seeks to fill the gap of scarce resources by introducing the first annotated dataset for Urdu medical NER. This contribution is expected to bridge the gap between NLP and healthcare in Urdu language, providing data and insights that would have a significant impact on clinical text processing.

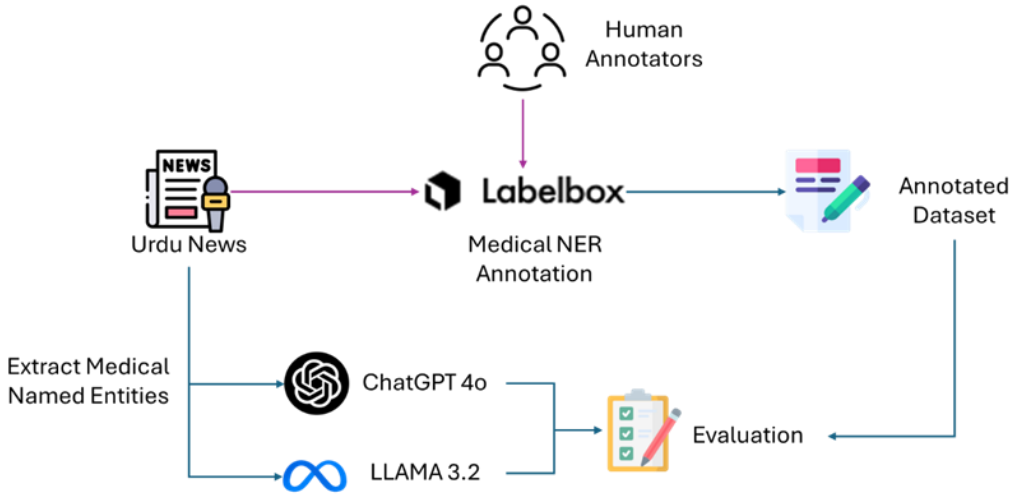

Figure 1: Workflow of Evaluating LLMs on Urdu Medical NER Task

3. Methodology

This section describes the dataset used and the large language models (LLMs) evaluated in this study for the medical named entity recognition (NER) task. We used a publicly available dataset of Urdu news[1], which includes news headlines and summaries. The subset of the complete dataset comprising of 2057 Urdu news headlines was carefully annotated for five medical named entities: *Disease, Symptom, Treatment, Medication, and Body Part*. The dataset was used to evaluate the performance of LLMs in extracting medical entities from the given Urdu text. Figure 1 illustrates the workflow of the methodology.

3.1 Dataset Preparation

The dataset of 2057 news summaries were annotated manually using LabelBox[2]. Four human annotators with educational background in medical sciences were recruited for NER task. The annotators were given a demonstration of the tool and were provided with the definition of each entity tag. The first 1000 news summaries were labeled independently by two annotators. F1-Score and percentage of exact agreement was calculated to assess the Inter-annotator agreement, yielding scores of **0.91** and **90.56%,** respectively. For the final entity labels, the discrepancies and conflicts were resolved through group discussions among the annotators and the research team. After confirming consistent annotation quality, the remaining 1057 headlines were labeled by a single annotator. Figure 2 presents a few examples of annotated news summaries.

3.2 LLMs Evaluated

In this study, we evaluated two large language models (LLMs) for the task of medical named entity recognition (NER) in Urdu: ChatGPT 4o and

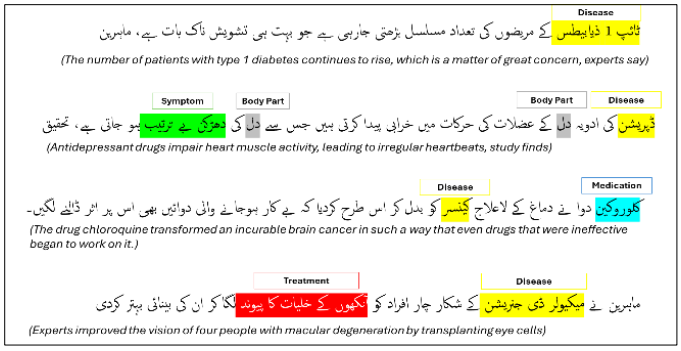

Figure 2: Instances of Annotated News Summaries

LLaMA 3.2. These models were selected due to their multilingual capabilities, accessibility, and increasing adoption in NLP research. Both models were evaluated in a zero-shot setting without task-specific fine-tuning.

3.2.1 ChatGPT-4o

ChatGPT-4o, developed by OpenAI, is a large-scale multimodal language model trained on diverse web-scale corpora. The model supports multiple languages, including Urdu. In this study, ChatGPT 4o was accessed through its web interface and prompted directly for entity extraction without additional fine-tuning or parameter adaptation. The evaluation therefore reflects its zero-shot performance on Urdu medical text.

3.2.2 LLaMA 3.2

LLaMA 3.2, developed by Meta, is an open-weight large language model designed for multilingual and instruction-following tasks. Similar to ChatGPT 4o, LLaMA 3.2 was evaluated in a zero-

1 https://github.com/mwaseemrandhawa/Urdu-News-Headline-Dataset

2 https://labelbox.com/

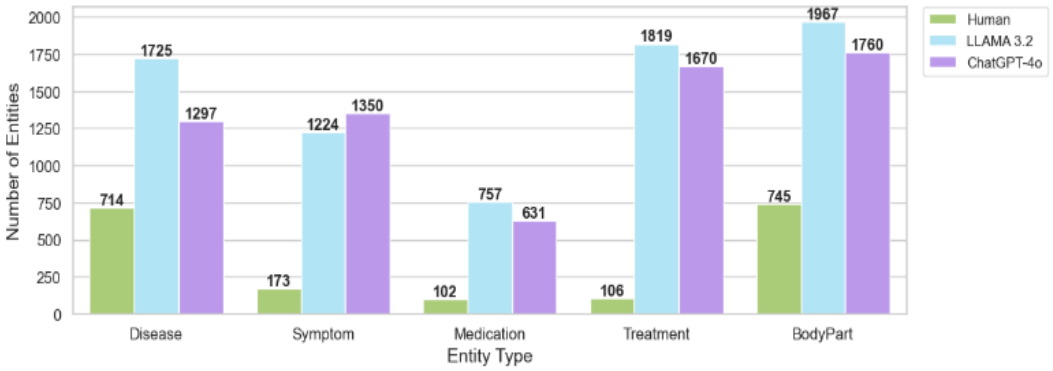

Figure 3: Entities Extracted by LLMs vs Humans

shot setting using its web-based interface. No domain-specific adaptation, fine-tuning, or parameter-efficient techniques were applied.

3.3 Prompt Design and Evaluation Procedure

To assess the capability of the LLMs for medical NER in Urdu, we designed a structured prompt specifying the five target entity categories: *Medication, Symptom, Disease, Treatment, and Body Part*. We refined prompt through preliminary experiments to ensure consistency in model outputs. The final prompt used for evaluation was:

Prompt
"*Extract the following information medication, symptom, disease, treatment, body part from given news summaries. Make a table with headers news, medication, symptom, disease, treatment, body part.*"

3.4 Evaluation Measures

To evaluate the performance of the LLMs on the medical NER task, we used standard entity-level precision, recall, and F1-score metrics. Evaluation was performed by comparing the entities extracted by the LLMs with the manually annotated gold-standard dataset.

We adopted an exact entity-level matching criterion. An extracted entity was considered a ***true positive (TP)*** only if the predicted entity text exactly matched the gold-standard entity span, and the predicted entity was assigned the correct entity category (e.g., Disease, Symptom, Medication, Treatment, or Body Part). If a model extracted an entity that did not appear in the gold-standard annotations, it was counted as a ***false positive (FP).*** If a gold-standard entity was not identified by the model, it was counted as a ***false negative (FN).***

For each entity category, precision, recall, and F1-score were computed. In addition to category level evaluation, we assessed the overall performance using micro-averaging across all entity categories.

4. Results

The annotated dataset contained a total of 2057 Urdu news headlines with brief summaries, categorized into five medical entity types: Disease, Symptom, Treatment, Medication, and Body Part. Table 1 summarizes the annotation statistics, including the total number of annotations and the count of unique entities present in each category.

Table 1: Summary Statistics of Entities in the Urdu News Dataset

Category	**# of Annotations**	**Unique Entities**
Disease	714	333
Symptom	173	129
Medication	102	75
Treatment	106	81
Body Part	745	226

We then analyzed the total number of entities extracted by the LLMs across the five medical categories. Figure 3 presents a comparison between human annotations, ChatGPT-4o, and LLAMA 3.2. The results highlight a notable difference in the count of entities which are extracted by human annotators and LLMs. Across all entity categories, both models generated substantially higher numbers of entity mentions compared to the gold-standard annotations.

For the Disease category, human annotators identified 714 entities, whereas LLAMA 3.2 extracted 1,725 entities and ChatGPT-4o extracted 1,297. In the Symptom category, humans annotated 173 entities, compared to 1,224 extracted by LLAMA 3.2 and 1,350 by ChatGPT-4o. A similar pattern was observed for Medication, where 102 entities were annotated by humans, while LLAMA 3.2 and ChatGPT-4o identified 757 and 631 entities, respectively. In the Treatment category, human annotators labeled

Table 2: Entity-level evaluation results (exact match) for LLAMA 3.2 and ChatGPT-4o on the Urdu medical NER dataset

Entity	Model	TP	FP	FN	Precision	Recall	F1
Disease	LLAMA 3.2	403	1322	311	0.234	0.564	0.33
	ChatGPT-4o	352	945	362	0.271	0.493	**0.35**
Symptom	LLAMA 3.2	73	1151	100	0.06	0.422	**0.105**
	ChatGPT-4o	71	1279	102	0.053	0.41	0.093
Medication	LLAMA 3.2	42	715	60	0.055	0.412	0.098
	ChatGPT-4o	36	595	66	0.057	0.353	0.098
Treatment	LLAMA 3.2	35	1784	71	0.019	0.33	**0.036**
	ChatGPT-4o	10	1660	96	0.006	0.094	0.011
Body Part	LLAMA 3.2	303	1664	442	0.154	0.407	0.223
	ChatGPT-4o	331	1429	414	0.188	0.444	**0.264**
Micro Avg	LLAMA 3.2	856	6636	984	0.114	0.465	0.183
	ChatGPT-4o	**800**	**5908**	**1040**	**0.119**	**0.435**	**0.187**

106 entities, whereas LLAMA 3.2 extracted 1,819 and ChatGPT-4o extracted 1,670 entities. Finally, in the Body Part category, humans annotated 745 entities, compared to 1,967 identified by LLAMA 3.2 and 1,760 by ChatGPT-4o. Overall, both LLMs produced substantially higher entity counts than the human-annotated gold standard.

4.1 Performance Evaluation of LLMs

We compared the performance of the LLMs using entity-level precision, recall, and F1-score against the manually annotated gold-standard labels. Table 2 presents the detailed evaluation results for both LLAMA 3.2 and ChatGPT-4o. Overall, the results indicate low precision but comparatively higher recall across most entity categories, indicating over-generation.

For the *Disease* category, ChatGPT-4o slightly outperformed LLAMA 3.2 in terms of F1-score, achieving 0.35 compared to 0.33. Although LLAMA 3.2 achieved higher recall (0.564 vs 0.493), ChatGPT-4o's higher precision (0.271 vs 0.234) contributed to its superior F1-score. In the *Body Part* category, ChatGPT-4o also outperformed LLAMA 3.2, achieving an F1-score of 0.264 compared to 0.223. ChatGPT-4o demonstrated both higher precision (0.188 vs 0.154) and recall (0.444 vs 0.407).

For *Symptom*, LLAMA 3.2 achieved a slightly higher F1-score (0.105) than ChatGPT-4o (0.093). Recall remained moderate for both models (0.422 for LLAMA 3.2 and 0.41 for ChatGPT-4o), but precision was very low (<0.06), indicating over-generation of entities.

In the *Medication* category, both models achieved similar performance (F1-scores of 0.098). LLAMA 3.2 had slightly higher recall (0.412 vs 0.353), whereas ChatGPT-4o had marginally better precision (0.057 vs 0.055). The *Treatment* category remained the most challenging, with low F1 scores for both models. LLAMA 3.2 achieved 0.036, while ChatGPT-4o only achieved 0.011. The particularly low recall of ChatGPT-4o (0.094) highlights difficulty in identifying treatment-related entities in Urdu medical text.

Micro-averaged performance shows that ChatGPT-4o slightly outperforms LLAMA 3.2 overall. ChatGPT-4o achieved a micro-average F1-score of 0.187 compared to 0.183 for LLAMA 3.2. Despite both models exhibiting low precision (0.119 vs 0.114), their moderate recall (0.435 for ChatGPT-4o and 0.465 for LLAMA 3.2) highlights

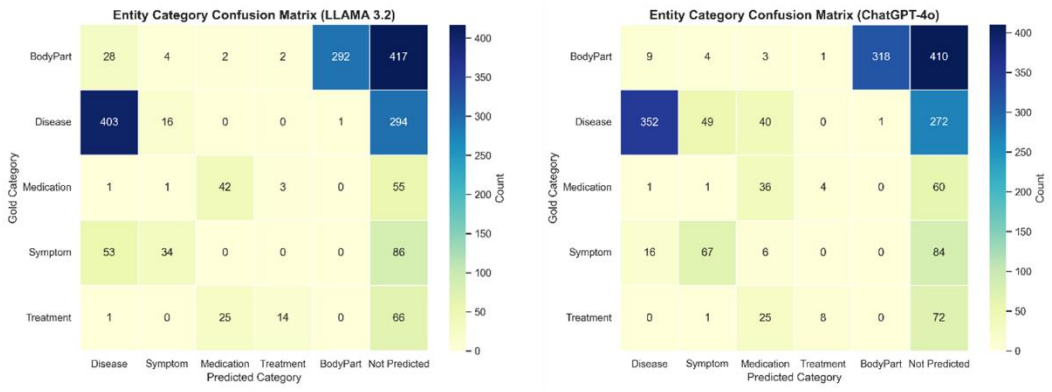

Figure 4: Confusion matrices comparing predicted versus gold-standard entity categories for LLAMA 3.2 and ChatGPT-4o on the Urdu medical NER dataset

the general tendency of LLMs to over-predict entities across all categories.

4.2 Error Analysis

We conducted an error analysis at the entity level to better understand the performance patterns of both LLMs. False positives (FPs) and false negatives (FNs) were examined. The analysis showed that following were the common types of mistakes made by LLMs in the Urdu medical domain:

4.2.1 Over generation of general terms (False Positives)

Both LLMs tend to over-generate general or common terms that are not specific entities in the gold standard. LLAMA 3.2 frequently predicted terms like "عام صحت" *(general health)* and " مختلف بیماریاں" (different diseases) as 'Disease', " علاج کرنا" (*do treatment*) as 'Treatment'. ChatGPT-4o over-predicted similar generic terms such as "مختلف امراض" *(various diseases*) as 'Disease' and "عمومی صحت" (*General Health*) as 'Medication'. These patterns highlight a tendency of both models to capture broadly related concepts rather than the precise entities defined in the annotated dataset.

4.2.2 Missed Critical Entities (False Negatives)

Both models sometimes failed to identify clinically important entities. LLAMA 3.2 missed key Disease entities like "کینسر" (*cancer*) and "موٹاپے" (*obesity*). ChatGPT-4o also missed similar entities in Disease, including "موٹاپے" (*obesity)*, "کینسر" (*cancer*) and "ذیابیطس" (*diabetes*). For Symptom, Medication, Treatment, and Body Part categories, both models missed essential entities such as "سردرد" (*headache, Symptom)*, *"* انسولین *"* *(insulin, Medication)*, "ہڈیوں" *(bones, Body Part).*

4.2.3 Morphological and Contextual Variation Challenges

Both LLMs struggled with handling morphological variations and context-specific forms in Urdu medical text. Several examples illustrate this limitation:

Pluralization issues: In a headline containing "دماغوں" (brains, plural), both LLAMA 3.2 and ChatGPT-4o extracted only "دماغ" (brain, singular), resulting in a mismatch with the gold-standard entity. Similarly, the plural "موٹاپے" (obesity, plural form) was normalized to "موٹاپا" (singular) by the models.

Compound and context-specific terms: In the phrase "دماغی خلیات" (brain cells), only "دماغ" was identified by the models, while the full entity was not captured.

4.2.4 Category Confusion Patterns

The confusion matrices shown in Figure 4 highlight systematic cross-category errors:

4.2.4.1 Body Part vs Disease/Symptom:
LLAMA 3.2 misclassified 28 Body Part entities as Disease, while ChatGPT-4o misclassified only 9.

4.2.4.2 Symptom vs Disease
Both models confused Symptom and Disease entities, with 53 (LLAMA 3.2) and 16 (ChatGPT-4o) Symptom entities predicted as Disease.

4.2.4.3 Treatment vs Medication
Treatment entities were often misclassified as Medication (14 instances for LLAMA 3.2, 25 for ChatGPT-4o).

These patterns indicate that both models not only over-generate general terms but also struggle to correctly assign entities to the appropriate category, particularly when categories share semantic similarity (e.g., Disease vs Symptom).

5. Discussion

This study evaluated the performance of two state-of-the-art LLMs, ChatGPT-4o and LLAMA 3.2, for medical NER in Urdu medical text. The findings highlight both the potential and the limitations of LLMs in handling low-resource languages like Urdu, particularly in the medical domain. While ChatGPT-4o achieved slightly better performance in terms of micro-averaged precision, recall, and F1-score (Micro F1: 0.187) compared to LLAMA 3.2 (Micro F1: 0.183), the

overall moderate scores for both models underscore the need for more specialized training and domain-specific optimization.

The annotation statistics provided valuable insights into model behavior. Both LLMs extracted substantially more entities than human annotators, often resulting in over-generation. This trend was particularly notable in the Body Part category, where ChatGPT-4o extracted 2,242 entities and LLAMA 3.2 extracted 2,105, compared to 745 annotated by humans. Such over-generation reflects the models' tendency to overgeneralize linguistic patterns learned during pretraining, leading to false positives, especially for generic or common terms. For instance, LLAMA 3.2 repeatedly predicted the term "جسم" (body) as a Body Part entity, occurring 796 times, while ChatGPT-4o predicted it 411 times. This illustrates the tendency of both models to over-generate common or generic terms, contributing significantly to false positives.

Beyond over-generation, both models exhibited challenges in handling morphological and variations. Cases such as "دماغوں" (brains) versus "دماغ" (brain), "دماغی خلیات" (brain cells) versus "دماغ", and "موٹاپے" versus "موٹاپا" demonstrate that LLMs often failed to fully capture plural forms, and compounds. This also highlights a language-specific challenge inherent to Urdu, which is characterized by rich morphology and complex word formations.

Performance varied across entity categories. ChatGPT-4o achieved higher F1-scores in Disease (0.35), Body Part (0.264), and slightly in Medication (0.098), while LLAMA 3.2 outperformed in Symptom (0.105). Both models struggled with Treatment, achieving F1-scores of 0.036 (LLAMA 3.2) and 0.011 (ChatGPT-4o). This could be attributed to the inherent ambiguity of treatment-related terminology in Urdu, which often involves multi-word expressions such as " گامانائف آئی کون مشین" (*Gama Knife Icon Machine*). These challenges are further exacerbated by the limited availability of pretraining data in the healthcare domain for Urdu, which restricts the models' ability to accurately disambiguate nuanced medical terms. A similar study (Naguib et al., 2024) also highlights these challenges where low-resource languages and domain-specific tasks showed suboptimal performance due to limited linguistic and domain coverage in pretraining corpora.

Overall, the results underscore the inherent challenges of medical NER in low-resource languages. Unlike high-resource languages such as English, which benefit from extensive annotated datasets and domain-specific corpora, Urdu suffers from limited labeled data, diverse linguistic structures, and terminological variations. These limitations contributed to low precision and moderate recall, even when models captured many entities. Nonetheless, the study demonstrates the feasibility of applying LLMs to medical NER in Urdu. ChatGPT-4o showed a slight edge in overall micro-averaged performance.

The performance of the LLMs could be potentially improved using fine tuning and few-shot learning techniques (Moscato et al., 2023). These techniques have shown promising improvements in the medical NER task in other languages (Rohanian et al., 2024; Zhu et al., 2024). Future work should focus on morphology-aware preprocessing to handle plurals and compound entities, domain-specific fine-tuning to reduce over-generation, and refined prompt engineering. These approaches, combined with few-shot and transfer learning, have potential for more accurate and scalable medical NER in Urdu and other low-resource languages.

6. Conclusion

This study evaluated the performance of two state-of-the-art LLMs, ChatGPT-4o and LLAMA 3.2, on medical NER in Urdu language. ChatGPT-4o generally outperformed LLAMA 3.2 across most metrics, highlighting its relative strength in extracting medical entities from low-resource language text. Despite this, both models demonstrated only moderate performance, reflecting challenges such as over-generation, difficulty with medical terminologies, and limited pretraining data in the Urdu medical domain. These findings underscore the need for targeted fine-tuning, morphology-aware preprocessing, and the development of larger, high-quality annotated datasets to improve LLM effectiveness. Overall, this study provides a foundational benchmark for applying NLP techniques to medical text in low-resource languages and informs future research aimed at enhancing LLM-based NER in similar contexts.

7. Acknowledgments

None.

8. Data Availability Statement

The data that support the findings of this study are available from the corresponding author (zarmeen.nasim@aku.edu) upon reasonable request.

9. Bibliographical References

Ahmed, A., Huang, D., & Arafat, S. Y. (2024). Enriching Urdu NER with BERT Embedding, Data Augmentation, and Hybrid Encoder-CNN Architecture. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *23*(4). https://doi.org/10.1145/3648362

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., &

McDermott, M. B. A. (2019). *Publicly Available Clinical BERT Embeddings*. https://arxiv.org/abs/1904.03323v3

Anam, R., Anwar, M. W., Jamal, M. H., Bajwa, U. I., De la Torre Diez, I., Alvarado, E. S., Flores, E. S., & Ashraf, I. (2024). A deep learning approach for Named Entity Recognition in Urdu language. *PLOS ONE*, *19*(3), e0300725. https://doi.org/10.1371/JOURNAL.PONE.0300725

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. El, Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., … Wu, Y. (2023). *PaLM 2 Technical Report*. https://arxiv.org/abs/2305.10403v3

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Coleman, J., Krishnamachari, B., Iskarous, K., & Rosales, R. (2024). *LLM-Assisted Rule Based Machine Translation for Low/No-Resource Languages*. 67–87. https://doi.org/10.18653/v1/2024.americasnlp-1.9

Deng, X., Bashlovkina, V., Han, F., Baumgartner, S., & Bendersky, M. (2023). LLMs to the Moon? Reddit Market Sentiment Analysis with Large Language Models. *ACM Web Conference 2023 - Companion of the World Wide Web Conference, WWW 2023*, 1014–1019. https://doi.org/10.1145/3543873.3587605

Elshin, D., Karpachev, N., Gruzdev, B., Golovanov, I., Ivanov, G., Antonov, A., Skachkov, N., Latypova, E., Layner, V., Enikeeva, E., Popov, D., Chekashev, A., Negodin, V., Frantsuzova, V., Chernyshev, A., & Denisov, K. (2024). *From General LLM to Translation: How We Dramatically Improve Translation Quality Using Human Evaluation Data for LLM Finetuning* (pp. 247–252). https://aclanthology.org/2024.wmt-1.17

Iwakami, Y., Takuma, H., & Iwashita, M. (2024). Complementary Method for NER by Extracting Proper Nouns from Images when Utilizing LLM. *2024 IEEE/ACIS 9th International Conference on Big Data, Cloud Computing, and Data Science (BCD)*, 233–238. https://doi.org/10.1109/BCD61269.2024.10743114

Jahan, I., Laskar, M. T. R., Peng, C., & Huang, J. X. (2024). A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, *171*. https://doi.org/10.1016/j.compbiomed.2024.108189

Kumichev, G., Blinov, P., Kuzkina, Y., Goncharov, V., Zubkova, G., Zenovkin, N., Goncharov, A., & Savchenko, A. (2024). MedSyn: LLM-Based Synthetic Medical Text Generation Framework. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *14950 LNAI*, 215–230. https://doi.org/10.1007/978-3-031-70381-2_14

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, 260–270. https://doi.org/10.18653/V1/N16-1030

Leaman, R., Khare, R., & Lu, Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, *57*, 28–37. https://doi.org/10.1016/J.JBI.2015.07.010

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020a). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/BIOINFORMATICS/BTZ682

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020b). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/BIOINFORMATICS/BTZ682

Moscato, V., Postiglione, M., & Sperlí, G. (2023). Few-shot Named Entity Recognition: Definition, Taxonomy and Research Directions. *ACM Transactions on Intelligent Systems and Technology*, *14*(5). https://doi.org/10.1145/3609483/ASSET/CD30433D-FB6D-4CEE-BDA4-86AC6442DA2D/ASSETS/GRAPHIC/TIST-2022-10-0375-F14.JPG

Naguib, M., Tannier, X., & Névéol, A. (2024). Few-shot clinical entity recognition in English, French and Spanish: masked language models outperform generative model prompting. *Proceedings of the 2nd. Clinical Natural Language Processing*

Workshop, 6829–6852.

Nasim, Z., Abidi, S., & Haider, S. (2020). Modeling POS Tagging for the Urdu Language. *2020 International Conference on Emerging Trends in Smart Technologies, ICETST 2020*. https://doi.org/10.1109/ICETST49965.2020.9080721

Nasim, Z., & Haider, S. (2022). Automatic Labeling of Clusters for a Low-Resource Urdu Language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *21*(5). https://doi.org/10.1145/3511097/ASSET/75F5299D-108C-4D86-919D-9CD617162B10/ASSETS/GRAPHIC/TALLIP-20-0576-INLINE10.GIF

Naz, S., Umar, A. I., Shirazi, S. H., Khan, S. A., Ahmed, I., & Khan, A. A. (2014). Challenges of Urdu named entity recognition: A scarce resourced language. *Research Journal of Applied Sciences, Engineering and Technology*, *8*(10), 1272–1278. https://doi.org/10.19026/RJASET.8.1095

Rohanian, O., Nouriborji, M., Kouchaki, S., Nooralahzadeh, F., Clifton, L., & Clifton, D. A. (2024). Exploring the effectiveness of instruction tuning in biomedical language processing. *Artificial Intelligence in Medicine*, *158*, 103007. https://doi.org/10.1016/J.ARTMED.2024.103007

Wang, Z., Zhou, Q., Junfeng, Z., Wang, Y., Ding, H., & Song, J. (2023). A Knowledge-Enhanced Medical Named Entity Recognition Method that Integrates Pre-Trained Language Models. *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, 296–301. https://doi.org/10.1109/MEDAI59581.2023.00046

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L., & Tang, Y. (2023). A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research*, *1*(6), 831–833. https://doi.org/10.1016/J.FMRE.2021.11.011

Zhao, Y., Wang, C., & Fu, G. (2012). A CRF Sequence Labeling Approach to Chinese Punctuation Prediction. *Pacific Asia Conference on Language, Information and Computation*.

Zhu, Z., Zhao, Q., Li, J., Ge, Y., Ding, X., Gu, T., Zou, J., Lv, S., Wang, S., & Yang, J. J. (2024). Comparative Analysis of Large Language Models in Chinese Medical Named Entity Recognition. *Bioengineering 2024, Vol. 11, Page 982*, *11*(10), 982. https://doi.org/10.3390/BIOENGINEERING11100982