KE-UMNER: Knowledge-Enriched Urdu Multimodal Named Entity Recognition Using LLM and Vision-Language Integration

Anonymous EMNLP submission

Abstract

Multimodal Named Entity Recognition (MNER) focuses on identifying entities of predefined categories within text by utilizing information from multiple sources, primarily text and images. While this task has seen progress in high-resource languages, it remains challenging for low-resource settings like Urdu, where social media content is often short, informal, and ambiguous. To address this, we propose KE-UMNER, a knowledge-enriched MNER framework that augments multimodal input with external semantic knowledge. It leverages Large Language Models to generate entity-specific contextual knowledge and employs a vision-language model (BLIP) to produce natural language captions from images. These knowledge signals are integrated with the input through a cross-modal attention mechanism and decoded via a BiLSTM-CRF layer for sequence labeling. Experiments on the Twitter2015-Urdu dataset show that KE-UMNER achieves a 12.08% absolute improvement in F1-score over prior state-of-the-art models. Ablation studies confirm the contribution of external knowledge sources, and case analyses demonstrate improved disambiguation in noisy, low-resource contexts.

1 Introduction

Named Entity Recognition (NER) is a fundamental task in natural language processing (Grishman and Sundheim, 1996) that involves identifying and classifying mentions of entities such as persons, organizations, and locations within unstructured text (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). While significant progress has been made in NER for high-resource languages, particularly under clean, monomodal conditions, the growing prevalence of multimodal content on social media platforms has introduced new challenges. Posts on platforms such as Twitter and Instagram often contain brief, informal textual content accompanied by images, creating a complex interaction between modalities that demands more sophisticated modeling.

(a) **Text: [Paula (MISC)]** has been waiting to be adopted for 2 years! Could U offer her a home?

(b) **Text: [Attenborough(PER)]** and **[Ben Kingsley(PER)]** with their **[Oscars(MISC)]**.

Figure 1: Two examples of Multimodal Named Entity Recognition. Named entities and their types are highlighted.

Multimodal Named Entity Recognition (MNER) addresses these challenges by leveraging both visual and textual content for more accurate entity classification (Mai et al., 2024; Yu et al., 2020). In this context, images serve as crucial disambiguation cues, especially when the textual context is insufficient or ambiguous (Liu et al., 2024a). As shown in Figure 1, traditional NER systems may misclassify entities due to ambiguity. For instance, in (a), the name “Paula” could be interpreted as a person, but the accompanying image reveals it refers to a dog, suggesting a MISC classification. Similarly, in (b), the term "Oscars" might be misinterpreted as an organization, but the image clarifies it refers to the prestigious film awards.

However, MNER becomes notably more challenging in low-resource languages like Urdu. Social media posts in Urdu are typically short, lack capitalization, and exhibit flexible word order and rich morphology, factors that amplify the ambiguity inherent in brief texts. These linguistic challenges are discussed in detail in Appendix A. Despite Urdu being spoken by over 70 million native and more than 100 million second-language speakers across

Pakistan and India, it remains underexplored in MNER compared to high-resource languages like English (Liang et al., 2024). The scarcity of annotated datasets and strong baselines further complicates model development and evaluation (Ahmed et al., 2024).

To address these challenges, we propose KE-UMNER, a Knowledge-Enriched Multimodal Named Entity Recognition framework. Rather than relying solely on limited in-text or in-image cues, KE-UMNER enriches multimodal inputs with external knowledge from vision-language models and large language models. Specifically, we utilize BLIP (Li et al., 2022) to generate image captions that translate visual content into natural language, and incorporate LLM-generated entity-specific prompts that provide supplementary context for ambiguous mentions. These signals are integrated into the model through a cross-modal fusion architecture followed by a structured prediction layer for sequence labeling.

We evaluate KE-UMNER on the Twitter2015-Urdu dataset (Ahmad et al., 2025), which reflects the noisy, multimodal nature of real-world social media in a low-resource setting. KE-UMNER significantly outperforms strong unimodal and multimodal baselines, particularly where text and image cues alone are insufficient. These results highlight the value of external knowledge as a complementary modality and suggest new directions for NER in low-resource scenarios.

The key contributions of our work are as follows:

- We introduce KE-UMNER, a knowledge-enriched framework for MNER in low-resource languages, which incorporates external knowledge to address ambiguity in short, informal social media text.
- Our approach fuses LLM-generated contextual cues and BLIP-derived image captions using a cross-modal attention mechanism, enabling effective integration of textual, visual, and external signals.
- We evaluate KE-UMNER on the Twitter2015-Urdu dataset, achieving significant improvements over unimodal and multimodal baselines, especially in ambiguous cases. Ablation and case studies further validate the contributions of each module in our framework.

2 Related Work

2.1 Multimodal Named Entity Recognition

Multimodal Named Entity Recognition extends traditional NER by incorporating visual context alongside text. Early NER systems relied solely on textual features, leveraging CNNs (Collobert et al., 2011) and BiLSTM-CRF architectures (Huang et al., 2015). The advent of pre-trained language models like BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020; Souza et al., 2020) led to significant improvements through contextualized textual representations.

Recent MNER methods focus on improving cross-modal alignment via cross-attention (Li et al., 2023; Yu et al., 2020), hierarchical visual prefixes (Chen et al., 2022b), and variational autoencoders (Zhou et al., 2022). Other enhancements include uncertainty modeling (Liu et al., 2022), relation classification (Sun et al., 2020; Xu et al., 2022), and query grounding (Jia et al., 2023). To enrich semantic representations, external knowledge is often introduced through scene graphs (Wang et al., 2023), multimodal graphs (Zhang et al., 2021), or label-based supervision (Wang et al., 2022).

However, the majority of MNER research has focused on high-resource languages. Low-resource languages like Urdu remain underexplored and present unique challenges due to scarce annotations, flexible word order, and complex morphology. These limitations highlight the need for more adaptable, knowledge-enhanced MNER frameworks capable of effectively handling noisy, multimodal data in under-resourced settings.

2.2 Pre-trained Vision and Language Models

Pre-trained vision-language models (VLMs) have advanced multimodal learning by aligning textual and visual features through large-scale image–text pretraining. Models such as ViLBERT (Lu et al., 2019), UNITER (Chen et al., 2020), and CLIP (Radford et al., 2021) provide strong multimodal representations transferable to downstream tasks. BLIP (Li et al., 2022) further improves alignment through bootstrapped image captioning, converting visual content into natural language for easier integration. LLMs like GPT offer complementary contextual knowledge, generating entity descriptions and background prompts that compensate for missing or ambiguous textual cues (Hou et al., 2024).

Our work integrates both modalities: BLIP-generated captions serve as visual knowledge, while GPT-derived prompts offer entity-level context. These signals are fused with the original input through a unified encoder, improving cross-modal alignment and enhancing robustness in low-resource scenarios.

3 Methodology

3.1 Task Formulation

We formulate MNER as a sequence labeling task. Given a text-image pair (T, V), where $T = \{w_1, w_2, \ldots, w_n\}$ is a sequence of n tokens and V is the associated image, the goal is to predict a label sequence $Y = \{y_1, y_2, \ldots, y_n\}$. Each y_i corresponds to the entity label for token w_i, following the BIO tagging scheme, where $y_i \in \{\text{B-type}, \text{I-type}, \text{O}\}$ and *type* denotes predefined entity categories such as `PER`, `LOC`, `ORG`, or `MISC`.

3.2 Overall Architecture

KE-UMNER is a knowledge-enriched MNER framework aimed at improving entity recognition in low-resource languages like Urdu. KE-UMNER integrates external knowledge from LLMs and VLMs to enhance entity disambiguation. The overall architecture, illustrated in Figure 2, consists of six modules: (1) LLM-Augmented Contextual Integration Module (LACIM), (2) Text Representation Module, (3) Visual Representation Module, (4) Text Self-Attention Module, (5) Cross-Modal Fusion Module, and (6) CRF Decoder Module.

LACIM enriches the input by retrieving entity-specific knowledge via GPT and generating semantic image captions through BLIP, forming a knowledge-enhanced sequence. The Text Representation Module processes this input using a pretrained Urdu-BERT, producing contextualized embeddings. Simultaneously, the Visual Representation Module extracts spatial features from images using ResNet-152, projecting them into a shared embedding space. The Text Self-Attention Module applies Transformer-based attention to highlight important contextual cues. The Cross-Modal Fusion Module (Yu et al., 2020) combines textual, visual, and external knowledge streams using multi-head attention and gating to prioritize relevant features. Finally, the CRF Decoder (Lafferty et al., 2001), comprising a BiLSTM and CRF layer, predicts BIO-tagged entity labels while modeling sequential dependencies. Together, these components enable KE-UMNER to robustly handle ambiguity and noise, delivering strong performance in challenging multimodal, low-resource settings.

3.3 LLM-Augmented Contextual Integration Module

LACIM is a core component of KE-UMNER that enhances entity recognition by enriching local textual and visual cues with external knowledge. Traditional MNER models often struggle with ambiguous mentions, especially in low-resource languages. To address this, LACIM integrates two knowledge extraction strategies: (1) LLM-based Contextual Knowledge Extraction (LCKE) and (2) BLIP-based Image Captioning with Urdu Translation.

The process involves two steps: First, entity candidates are identified, and relevant contextual knowledge is retrieved using a large language model (GPT). Second, image captions are generated via BLIP and translated into Urdu to align visual information with the text. This knowledge-enhanced input improves entity disambiguation, particularly when local context is insufficient.

3.3.1 LLM-Based Contextual Knowledge Extraction

LCKE enhances entity recognition by integrating external knowledge to improve disambiguation, particularly in low-resource languages like Urdu. Traditional NER models often struggle due to limited context; LCKE addresses this by retrieving rich, linguistically and culturally appropriate information using LLMs.

Identifying Entity Candidates. The first step in knowledge extraction is detecting potential entity mentions in the input text. KE-UMNER utilizes a Transformer-based encoder to process the sentence:

$$T_{\text{orig}} = (w_1, w_2, \ldots, w_n), \tag{1}$$

the model applies BIO tagging to classify each token. Entity spans are extracted as

$$S = \{s_1, s_2, \ldots, s_m\}, \tag{2}$$

where each s_i represents a detected entity mention. These spans are used as queries for knowledge retrieval.

Contextual Knowledge Extraction Using LLM. After detecting entity candidates, LCKE constructs structured prompts and queries GPT-3.5 to retrieve relevant contextual knowledge:

$$K_{\text{GPT}} = (k_1, k_2, \ldots, k_m). \tag{3}$$

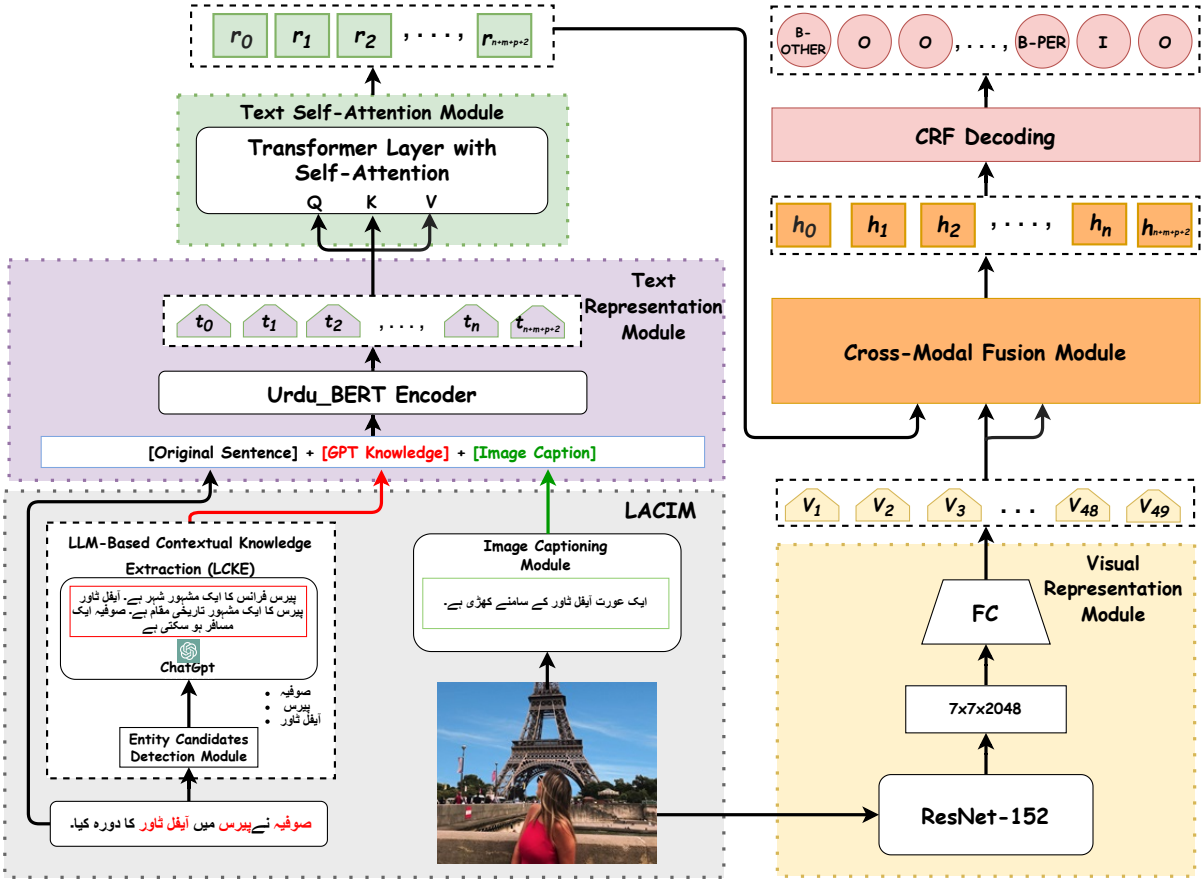

Figure 2: Overall Architecture of KE-UMNER Framework.

This external knowledge enriches the input representation, enhances semantic understanding, and improves entity disambiguation, particularly for low-resource languages like Urdu where annotated corpora are limited. Prompt construction is tailored according to entity type: biographical details for PER, historical and geographical context for LOC, organizational attributes for ORG, and conceptual information for MISC. Further details, including structured prompt templates and examples, are provided in Appendix B.

3.3.2 BLIP-Based Image Captioning

In multimodal settings, images can provide valuable cues for entity recognition, but not all images are directly helpful. To ensure effective utilization, we employ BLIP to generate descriptive captions that translate visual content into text.

Given an input image V, BLIP produces a caption:

$$C_{\text{IMG}} = (c_1, c_2, \ldots, c_p) \tag{4}$$

where c_i are the caption tokens. This caption is further translated into Urdu using GPT:

$$C_{\text{IMG\_Urdu}} = (c'_1, c'_2, \ldots, c'_p) \tag{5}$$

The LLM-extracted knowledge K_{GPT} and Urdu-translated captions $C_{\text{IMG\_Urdu}}$ are concatenated with the original input text to form a knowledge-enriched sequence. This augmented input is then processed by the Text Representation Module using Urdu-BERT.

By integrating structured knowledge and visual descriptions, KE-UMNER enhances disambiguation, aligns cross-modal information, and improves recognition performance in noisy Urdu social media content.

3.4 Text Representation Module

The Text Representation Module encodes knowledge-augmented inputs into contextual embeddings. It integrates the original text T_{orig}, LLM-extracted knowledge K_{GPT}, and BLIP-generated Urdu captions $C_{\text{IMG\_Urdu}}$, combined as:

$$T_{\text{final}} = [T_{\text{orig}}; K_{\text{GPT}}; C_{\text{IMG\_Urdu}}], \tag{6}$$

where each component is tokenized into $(w_1, \ldots, w_n, k_1, \ldots, k_m, c_1, \ldots, c_p)$. Following BERT conventions, we insert [CLS] and [SEP] tokens, and obtain input embeddings by summing

word, segment, and positional embeddings. Passing the resulting sequence through Urdu-BERT yields contextualized token representations:

$$T = (t_0, t_1, \ldots, t_{n+m+p+1}), \tag{7}$$

where each $t_i \in \mathbb{R}^d$ captures semantic and external knowledge features.

3.5 Visual Representation Module

The Visual Representation Module extracts structured visual features aligned with the textual modality. Each input image is resized to 224×224 pixels and passed through ResNet-152, producing a 7×7 grid of spatial features, each a 2048-dimensional vector:

$$U = \{u_1, u_2, \ldots, u_{49}\}. \tag{8}$$

To align with text embeddings, visual features are projected into a shared d-dimensional space via a linear transformation:

$$V = W_u^\top U = \{v_1, v_2, \ldots, v_{49}\}, \tag{9}$$

where $W_u \in \mathbb{R}^{2048\times d}$ is a learnable matrix. The resulting embeddings V are passed to the Cross-Modal Fusion Module for joint multimodal processing.

3.6 Text Self-Attention Module

To capture linguistic dependencies and contextual relationships in Urdu text, KE-UMNER incorporates a Transformer-based Self-Attention Layer. This module allows the model to dynamically weigh the importance of different tokens, crucial for disambiguating entities based on surrounding context.

Operating on the contextualized embeddings from Urdu-BERT, the input sequence is:

$$C = (c_0, c_1, \ldots, c_{n+1}) \tag{10}$$

The self-attention mechanism computes interactions among tokens:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{11}$$

where Q, K, and V are linear projections of C, and d_k is the dimensionality scaling factor.

The attention-weighted outputs are processed through a feed-forward network, producing the enhanced textual representations:

$$R = (r_0, r_1, \ldots, r_{n+1}) \tag{12}$$

These refined embeddings are then passed to the Cross-Modal Fusion Module.

3.7 Cross-Modal Fusion Module

To achieve cross-modal alignment, we adopt the Cross-Modal Fusion strategy based on the approach outlined in (Yu et al., 2020), adapting it to integrate textual, visual, and external knowledge representations. KE-UMNER uses a Cross-Modal Transformer (CMT) where visual features V are treated as queries and textual embeddings R as keys and values. Cross-modal attention is computed as:

$$\text{Attn}_i(V, R) = \text{softmax}\left(\frac{(W_i^Q V)^\top (W_i^K R)}{\sqrt{d/m}}\right)(W_i^V R)^\top \tag{13}$$

$$\text{MH-Attn}(V, R) = W'\Big([\text{CA}_1(V, R), \ldots, \text{CA}_m(V, R)]\Big)^\top \tag{14}$$

where W_i^Q, W_i^K, and W_i^V are learnable projection matrices across m attention heads. The fused features are refined via residual connections, Layer Normalization, and a Feed-Forward Network (FFN):

$$\tilde{P} = \text{LayerNorm}(V + \text{MH-CA}(V, R)), \tag{15}$$

$$P = \text{LayerNorm}(\tilde{P} + \text{FFN}(\tilde{P})), \tag{16}$$

producing multimodal embeddings P. To further refine alignment, a second CMT layer reverses the attention direction, using P as both queries and keys/values, resulting in image-aware word representations:

$$A = (a_0, a_1, \ldots, a_{n+1}), \tag{17}$$

where each a_i captures combined textual and visual semantics. Word-aware visual features are computed by querying visual regions with text embeddings:

$$Q = (q_0, q_1, \ldots, q_{n+1}), \tag{18}$$

where each q_i represents the most semantically aligned visual context for the corresponding token. A gating mechanism selectively integrates visual cues:

$$g = \sigma(W_a^\top A + W_q^\top Q), \tag{19}$$

where W_a and W_q are trainable parameters and σ is the sigmoid activation. The final word-aware visual representation is obtained as:

$$B = g \odot Q, \tag{20}$$

where $\odot$ denotes element-wise multiplication. This selective fusion emphasizes meaningful visual-textual interactions while mitigating irrelevant visual noise.

3.8 CRF Decoder Module

After obtaining the image-aware word representations A and word-aware visual representations B, we concatenate them to form the multimodal hidden states:

$$h_i = \text{Concat}(a_i, b_i), \quad \text{for } i = 0, \ldots, n+1, \quad (21)$$

resulting in $H = \{h_0, h_1, \ldots, h_{n+1}\}$, where each $h_i \in \mathbb{R}^{2d}$ integrates both textual and visual information.

The sequence H is then passed through a Bidirectional Long Short-Term Memory (BiLSTM) network to capture contextual dependencies from both directions. Each output h_i^{BiLSTM} is the concatenation of forward and backward hidden states, providing a richer representation for sequence labeling.

On top of the BiLSTM outputs, a Conditional Random Field (CRF) layer models dependencies between adjacent labels, improving sequence-level predictions. The probability of a label sequence $Y = \{y_0, y_1, \ldots, y_{n+1}\}$ is computed as:

$$P(Y \mid (T, V)) = \frac{\exp(\text{score}(H, Y))}{\sum_{Y' \in \mathcal{Y}} \exp(\text{score}(H, Y'))}, \quad (22)$$

where the score combines transition and emission scores across the sequence. Emission scores are computed by projecting each h_i^{BiLSTM} with a learned label-specific weight.

The model is trained by minimizing the negative log-likelihood of the correct label sequence:

$$\mathcal{L} = -\log(P(Y \mid (T, V))). \quad (23)$$

This objective encourages correct entity predictions while capturing global sequence consistency.

4 Experiments

4.1 Settings

Dataset We evaluate KE-UMNER on the Twitter2015-Urdu dataset, a multilingual extension of the widely used Twitter2015 benchmark. The dataset consists of short Urdu tweets paired with images, reflecting the informal and diverse nature of social media content. Each instance is annotated using the BIO tagging scheme with four entity types: `PER`, `LOC`, `ORG`, and `MISC`. Further dataset statistics are provided in Appendix C.

Baselines We compare KE-UMNER against two baseline categories: (1) Text-only models, including LSTM-CRF, BiLSTM-CRF (Huang et al., 2015), HBiLSTM-CRF (Lample et al., 2016), CNN-BiLSTM-CRF (Ma and Hovy, 2016), BERT (Devlin et al., 2019), and BERT-CRF (Souza et al., 2020), which rely solely on text and often struggle in visually rich contexts. (2) Multimodal models, such as UMT (Yu et al., 2020), RpBERT (Sun et al., 2021), HVPNET (Chen et al., 2022a), MAF (Xu et al., 2022), MGCMT (Liu et al., 2024b), and U-MNER(Ahmad et al., 2025), which jointly leverage text and image features for entity recognition. More details are provided in Appendix D.

Evaluation Metrics We evaluate model performance using standard NER metrics: Precision, Recall, and F1-score, with F1 as the primary metric. Precision reflects the correctness of predicted entities, while recall measures coverage of true entities. Metrics are computed using exact span matching, reported per entity type and overall.

Implementations Experiments were performed using PyTorch on a system with an NVIDIA RTX 4090 GPU. The Urdu-BERT base model was fine-tuned for textual embeddings, while ResNet-152 was used for visual feature extraction. Knowledge augmentation was done using GPT-3.5, and image captions were generated with the BLIP model (Li et al., 2022). Key hyperparameters included a learning rate of 5×10^{-5}, a dropout rate of 0.1, and a batch size of 16. The model processed sequences with a maximum length of 128 tokens and utilized 12 cross-modal attention heads for multimodal integration.

4.2 Results and Analysis

Table 1 presents the experimental results on the Twitter2015-Urdu dataset. We evaluate six text-only models and four multimodal models, reporting Precision (P), Recall (R), and F1-score (F1) across entity types: `PER` (Person), `LOC` (Location), `ORG` (Organization), and `MISC` (Miscellaneous).

First, we examine the performance of text-only models, including LSTM-CRF, BiLSTM-CRF, CNN-BiLSTM-CRF, BERT, and BERT-CRF. While LSTM-based models demonstrate reasonable results, transformer-based models such as BERT and BERT-CRF outperform them by leveraging stronger contextual representations. However, despite their improvements, text-only models struggle to resolve entity ambiguities in short, noisy, or visually grounded social media posts, where textual cues alone are insufficient.

Modality	Methods	Single Type (F1)				Overall		
		PER	LOC	ORG	MISC	P	R	F1
Text	LSTM-CRF	63.36	56.90	34.85	16.42	56.44	45.46	50.36
	BiLSTM-CRF	64.97	57.82	34.28	19.99	58.08	46.95	51.93
	HBiLSTM-CRF	67.83	58.71	30.92	18.05	57.10	49.04	52.76
	CNN-BiLSTM-CRF	58.23	48.51	28.61	19.47	56.96	45.23	50.42
	BERT	72.01	67.74	39.89	15.36	62.08	57.22	59.55
	BERT-CRF	72.85	68.16	41.29	16.65	63.32	56.87	59.92
Text+Vision	MGCMT	69.71	65.31	42.60	24.36	57.28	55.57	56.41
	MAF	73.75	69.32	47.09	25.30	59.85	61.30	60.57
	RpBERT	71.47	68.48	39.07	11.11	64.32	57.69	60.82
	UMT	68.08	64.77	40.71	25.80	58.21	55.41	54.07
	HVPNET	72.28	63.47	43.47	25.80	62.93	59.03	60.92
	U-MNER	73.83	70.71	47.91	23.32	63.27	62.24	62.75
	KE-UMNER (Ours)	**86.42**	**80.59**	**68.25**	**52.51**	**74.08**	**75.60**	**74.83**

Table 1: Performance comparison of text-based and multimodal NER models on the Twitter15-Urdu dataset. KE-UMNER achieves the best performance across all metrics.

Next, multimodal models, including UMT, RpBERT, HVPNET, and U-MNER, show consistent gains over text-only baselines. By incorporating visual information, these models better handle ambiguous or underspecified entities. Nevertheless, existing multimodal approaches are limited when image cues are weak or misleading, as they lack external semantic reinforcement.

Finally, KE-UMNER achieves the best overall performance, with an F1-score of 74.83%, significantly outperforming U-MNER (62.75%) by 12.08%. KE-UMNER also achieves strong per-entity F1-scores: 86.42% (PER), 80.59% (LOC), 68.25% (ORG), and 52.51% (MISC), maintaining balanced Precision (74.08%) and Recall (75.60%). These results demonstrate that integrating LLM-extracted knowledge and BLIP-generated captions effectively enhances multimodal entity recognition, particularly by mitigating ambiguities and enriching contextual understanding. By fusing textual, visual, and external knowledge signals, KE-UMNER sets a new benchmark for Urdu MNER in low-resource, multimodal settings.

4.3 Ablation Study

We evaluate the impact of KE-UMNER's knowledge components by ablating the LLM-based Contextual Knowledge Extraction (LCKE) and BLIP-generated image captions. Table 2 reports changes in Precision (P), Recall (R), and F1-score (F1). Removing LCKE leads to a significant F1 drop from 74.83% to 58.50%, underscoring its key role in resolving ambiguous mentions. Excluding image captions causes a smaller decline to 74.15%, showing that visual context is helpful but less essential. Removing both components drops F1 to 58.55%. These results highlight that KE-UMNER's strong performance stems from the synergy of multimodal inputs and external knowledge, with LLM-based knowledge playing the most significant role.

Methods	P	R	F1
KE-UMNER (Full Model)	**74.08**	**75.60**	**74.83**
w/o LCKE	57.38	59.57	58.50
w/o Image Captions	73.42	74.89	74.15
w/o LCKE & Image Captions	57.33	59.82	58.55

Table 2: Ablation results showing the effect of removing LLM-based knowledge (LCKE) and BLIP-generated image captions on KE-UMNER's performance.

4.4 Case Study

We highlight KE-UMNER's effectiveness using examples from the Twitter2015-Urdu dataset (Figure 3). Text-only models (e.g., BERT-CRF) and multimodal models (e.g., HVPNET, U-MNER) often misclassify ambiguous mentions, especially when visual cues are limited. KE-UMNER addresses this by combining text, BLIP captions, and LLM-based contextual knowledge.

Figure 3: Case study illustrating how LLM-extracted knowledge and image captions enhance entity recognition by resolving ambiguities in multimodal NER.

In case (a), "فیڈرر" (Federer) and "ومبلڈن" (Wimbledon) are mentioned. While Federer is correctly identified, Wimbledon is misclassified as LOC by baselines. KE-UMNER correctly labels it as MISC using LLM knowledge linking it to a tennis tournament.

In case (b), "وینس" (Venus) is misclassified due to vague context. Visual cues ("a woman in sports attire") offer limited help. KE-UMNER correctly predicts PER by recognizing Venus as a tennis player via LLM knowledge. These examples underscore the importance of integrating external knowledge with multimodal signals for disambiguating complex entity mentions.

5 Conclusion

We proposed KE-UMNER, a knowledge-enriched multimodal named entity recognition framework tailored for the Urdu language. By integrating LLM-generated prompts and BLIP-based image captions into a unified cross-modal architecture, KE-UMNER effectively enhances entity disambiguation in noisy, low-resource social media contexts. Experiments on the Twitter2015-Urdu dataset show that KE-UMNER significantly outperforms strong unimodal and multimodal baselines, particularly for ambiguous or visually grounded mentions. Our findings highlight the importance of external knowledge in improving MNER performance for under-resourced languages.

Limitations

While KE-UMNER demonstrates strong performance in low-resource multimodal NER, several limitations remain. First, the model relies on large-scale pretrained language and vision-language models, posing scalability challenges in resource-constrained environments. Second, the current study focuses exclusively on Urdu; the model's generalizability to other low-resource or morphologically rich languages is yet to be examined. Third, external knowledge sources, LLM-generated prompts and BLIP-based captions, are not guaranteed to be accurate and may introduce noise or

cultural bias. Addressing these limitations in future work involves exploring lightweight alternatives to current architectures, extending the framework cross-linguistically, and investigating semi-supervised approaches to mitigate reliance on annotated multimodal data.

Ethics Statement

This work aligns with the ACL Ethics Policy and aims to support linguistic equity by advancing NLP for low-resource languages. We use publicly available datasets and non-sensitive content. While external models like LLMs and VLMs enhance contextual understanding, they may introduce biases or inaccuracies. Future research should address these concerns to ensure fairness and cultural relevance in multimodal NER systems.

References

Hussain Ahmad, Qingyang Zeng, and Jing Wan. 2025. A benchmark dataset and a framework for urdu multimodal named entity recognition. *arXiv preprint arXiv:2505.05148*.

Anil Ahmed, Degen Huang, Syed Yasser Arafat, and Imran Hameed. 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):38.

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022a. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1607–1618, Seattle, United States. Association for Computational Linguistics.

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022b. Good visual guidance makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1607–1618, Seattle, United States. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *Computer Vision – ECCV 2020*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer, Cham.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537. J. Mach. Learn. Res. 12, null (2/1/2011).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Chenyu Hou, Gaoxia Zhu, Juan Zheng, Lishan Zhang, Xiaoshan Huang, Tianlong Zhong, Shan Li, Hanxiang Du, and Chin Lee Ker. 2024. Prompt-based and fine-tuned GPT models for context-dependent and -independent deductive coding in social annotation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 518–528, Kyoto, Japan. Association for Computing Machinery.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

M. Jia, L. Shen, X. Shen, L. Liao, M. Chen, X. He, Z. Chen, and J. Li. 2023. Mnerqg: An end-to-end mrc framework for multimodal named entity recognition with query grounding. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Xiujiao Li, Guanglu Sun, and Xinyu Liu. 2023. ESPVR: Entity spans position visual regions for multimodal named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7785–7794, Singapore. Association for Computational Linguistics.

X. Liang, R. Mao, L. Wu, J. Li, M. Zhang, and Q. Li. 2024. Enhancing low-resource nlp by consistency training with data and model perturbations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:189–199.

C. Liu, D. Yang, B. Yu, and L. Bu. 2024a. Dghc: A hybrid algorithm for multi-modal named entity recognition using dynamic gating and correlation coefficients with visual enhancements. *IEEE Access*, 12:69151–69162.

Luping Liu, Meiling Wang, Mozhi Zhang, Linbo Qing, and Xiaohai He. 2022. Uamner: Uncertainty-aware multimodal named entity recognition in social media posts. *Applied Intelligence*, 52:4109–4125.

Peipei Liu, Gaosheng Wang, Hong Li, Jie Liu, Yimo Ren, Hongsong Zhu, and Limin Sun. 2024b. Multi-granularity cross-modal representation learning for named entity recognition on social media. *Information Processing & Management*, 61(1):103546.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

W. Mai, Z. Zhang, K. Li, Y. Xue, and F. Li. 2024. Dynamic graph construction framework for multimodal named entity recognition in social media. *IEEE Transactions on Computational Social Systems*, 11(2):2513–2522.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, pages 403–417, Berlin, Heidelberg. Springer-Verlag.

Lin Sun, Jiquan Wang, Yindu Su, Fangsheng Weng, Yuxuan Sun, Zengwei Zheng, and Yuanyi Chen. 2020. RIVA: A pre-trained tweet multimodal model based on text-image relation for multimodal NER. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1852–1862, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: A text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13860–13868. AAAI Press.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

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

J. Wang, Y. Yang, K. Liu, Z. Zhu, and X. Liu. 2023. M3s: Scene graph driven multi-granularity multi-task learning for multi-modal ner. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:111–120.

Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. 2022. ITA: Image-text alignments for multi-modal named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3176–3189, Seattle, United States. Association for Computational Linguistics.

Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022. Maf: A general matching and alignment framework for multimodal named entity recognition. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22)*, pages 1215–1223, New York, NY, USA. Association for Computing Machinery.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages

3342–3352, Online. Association for Computational Linguistics.

D. Zhang, S. Wei, S. Li, H. Wu, Q. Zhu, and G. Zhou. 2021. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14347–14355.

Baohang Zhou, Ying Zhang, Kehui Song, Wenya Guo, Guoqing Zhao, Hongbin Wang, and Xiaojie Yuan. 2022. A span-based multimodal variational autoencoder for semi-supervised multimodal named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6293–6302, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Appendix

Entity Type	Train	Val	Test
PER	2,255	558	1,939
LOC	2,076	529	1,781
ORG	897	240	825
MISC	946	226	673
Total Entities	6,174	1,553	5,218
Total Tweets	4,000	1,000	3,257

Table 3: Entity and tweet distribution in the Twitter2015-Urdu dataset across train, validation, and test splits.

A Challenges in Urdu Named Entity Recognition

Urdu presents several unique challenges for NER, primarily due to its linguistic structure, informal usage, and lack of orthographic cues. Unlike English, Urdu lacks capitalization, making it difficult to distinguish proper nouns such as "پروین شاکر" (Parveen Shakir) from common words based only on surface form. Entity interpretation in Urdu is highly context sensitive. A token like "اسمرا" (Asmara) may denote a person in one sentence and a location in another, depending entirely on context. Ambiguity further arises from overlapping surface forms, such as "ریاض" (Riaz) being interpretable as either a person or the city "ریاض" (Riyadh). Morphological complexity adds another layer of difficulty. Inflectional and agglutinative processes can shift a named entity into a non-entity role. For instance, "نثار" (Nisar) becomes an adjective in "جان نثار" (self-sacrificing), and "مختار" (Mukhtar) becomes a descriptor in "خود مختار" (independent). Urdu also allows flexible word order, enabling multiple valid sentence structures with the same meaning. For example, "میں نے چترال میں ایک نیا ریسٹورنٹ کھولا۔" and "میں نے ایک نیا ریسٹوران چترال میں کھولا۔" express the same idea with different token sequences, challenging models that rely on positional patterns. Finally, spelling variations and orthographic inconsistencies are common in informal settings. The name "Samad" may appear as "صمد" "سمد" or "سماد" . Similarly, expressions like "کے ساتھ" (with) may appear joined as "کیساتھ" complicating tokenization. Figure 5 provides examples illustrating each of these challenges in detail.

B Prompt Construction for LLM-Based Contextual Knowledge Extraction

To retrieve relevant contextual knowledge for each entity, we design structured prompts tailored to the entity type. The prompt structure is adapted for four categories: `PER` (Person), `LOC` (Location), `ORG` (Organization), and `MISC` (Miscellaneous).

For `PER` entities, prompts focus on biographical details, achievements, and contributions. For `LOC` entities, the prompts emphasize historical significance, geography, and cultural relevance. For `ORG` entities, they retrieve information about the organization's mission, operations, and societal role. For `MISC` entities, the prompts aim to clarify concepts, events, or notable phenomena relevant to Urdu linguistic and cultural contexts.

Figure 4 illustrates the structured prompt construction process, outlining task descriptions, instruction templates, entity-specific examples, and expected outputs. This ensures that the LLM-generated knowledge is precise, culturally aligned, and supports accurate entity disambiguation.

C Dataset Details

The Twitter2015-Urdu dataset consists of 8,257 text-image pairs, specifically designed for Multimodal Named Entity Recognition in Urdu. It contains four entity types: `PER` (person), `LOC` (location), `ORG` (organization), and `MISC` (miscellaneous). The dataset is divided into training (80%), validation (10%), and test (10%) sets. The entity distribution is as follows: 2,255 person (`PER`) entities in training, 558 in validation, and 1,939 in the test set; 2,076 location (`LOC`) entities in training, 529 in validation, and 1,781 in the test set; 897 organization (`ORG`) entities in training, 240 in validation, and 825 in the test set; and 946 miscellaneous (`MISC`) entities in training, 226 in validation, and 673 in the test set. The distribution of entities across the dataset is summarized in Table 3.

<Task Descriptions>

You are an expert linguist specializing in Urdu. Your task is to generate a concise, meaningful, and culturally accurate context for the given Urdu entity. The context should align strictly with the entity's type (e.g., Person, Location, Organization, or Other). You must consider the linguistic, cultural, and historical significance of the entity within the context of Urdu-speaking communities, ensuring that the generated description reflects its relevance and importance in these communities.

<Instructions Based on Entity Type>

- For '**PER**' (Person): Focus on key achievements, cultural impact, and the individual's legacy within Urdu speaking communities.
- For '**LOC**' (Location): Emphasize the historical, geographical, and cultural significance, including landmarks, events, and local identity.
- For '**ORG**' (Organization): Provide insights into the organization's role in the community, its mission, and impact within a societal, economic, or cultural context.
- For '**OTHER**' (Other): Describe unique aspects of the entity, whether it's a concept, phenomenon, object, or event, highlighting its relevance to the Urdu-speaking world.

<Examples>

- **Entity:** آیفل ٹاور (Type: LOC)
  Context: آیفل ٹاور پیرس، فرانس کا مشہور لوہے کا مینار ہے جو دنیا بھر میں سیاحوں کی توجہ کا مرکز ہے۔
- **Entity:** البرٹ آئن سٹائن (Type: PER)
  Context: البرٹ آئن سٹائن ایک معروف سائنسدان ہیں جنہوں نے نظریہ اضافیت پیش کیا اور دنیا کی فزکس کو نیا رخ دیا۔
- **Entity:** گوگل (Type: ORG)
  Context: گل ایک عالمی سطح پر مشہور کمپنی ہے جو انٹرنیٹ سے متعلق مختلف خدمات فراہم کرتی ہے، جن میں سرچ انجن، ای میل، اور ایڈورٹائزنگ سروسز شامل ہیں۔
- **Entity:** بلیک ہول (Type: OTHER)
  Context: بلیک ہول ایک فلکیاتی مظہر ہے جس میں کشش ثقل اتنی زیادہ ہوتی ہے کہ اس سے روشنی بھی باہر نہیں نکل سکتی۔

<Expected Output Format>

Provide the context in the following format:

- **Context**: <**Urdu Context**>

Figure 4: Structured Prompt Construction for LLM-Based Contextual Knowledge Extraction.

D Baselines

To evaluate KE-UMNER, we compare it with two categories of baseline models: text-only models and multimodal models. Text-only models, such as LSTM-CRF, BiLSTM-CRF (Huang et al., 2015), HBiLSTM-CRF (Lample et al., 2016), CNN-BiLSTM-CRF (Ma and Hovy, 2016), BERT (Devlin et al., 2019), and BERT-CRF (Souza et al., 2020), rely solely on textual information for entity recognition. These models excel when the text is well-structured but struggle with ambiguous or visually dependent entities, which are common in social media posts. For instance, LSTM-CRF and its variants, such as BiLSTM-CRF, capture sequential dependencies but lack the ability to handle visual context, making them less effective for recognizing entities in noisy or multimodal settings. BERT and BERT-CRF, while improving upon these methods by providing context-aware embeddings, still rely entirely on textual data and miss the potential insights offered by accompanying images.

On the other hand, multimodal models, including UMT (Yu et al., 2020), RpBERT (Sun et al., 2021), HVPNET (Chen et al., 2022a), MAF (Xu et al., 2022), MGCMT (Liu et al., 2024b), and U-MNER(Ahmad et al., 2025), integrate both textual and visual features for enhanced entity recognition. These models utilize visual cues to disambiguate entities that may be difficult to classify based on text alone. For example, UMT and RpBERT use cross-modal attention mechanisms to align text and image representations, while HVPNET and MGCMT refine multimodal integration by processing image and text features at different levels of abstraction. However, these models often struggle to fully capture external contextual knowledge, limiting their ability to handle complex entities or resolve ambiguities effectively. U-MNER, the predecessor of KE-UMNER, represents an earlier approach to multimodal entity recognition but lacks the integration of external knowledge, which is a key feature of KE-UMNER's design. This comparison highlights KE-UMNER's strength in combining both multimodal cues and external knowledge for more robust entity recognition, particularly in complex, real-world social media contexts.

Challenges	Example Sentences
Absence of Capitalization	میری ملاقات پروین شاکر سے ہوئی۔ ("I met Parveen Shakir.") → "Parveen Shakir" as <Person>
Context Sensitivity	اسمرا بہت خوبصورت ہے۔ ("Asmara is very beautiful.") → "Asmara" is interpreted as a <Person> اریٹیریا کی آزادی کے بعد اسمرا نے تیزی سے ترقی کی۔ (Asmara experienced rapid development after Eritrea's independence.) → "Asmara" is interpreted as a <Location>
Ambiguous Named Entities	ریاض سعودی عرب کے شہر ریاض میں اپنے دوست کے ساتھ مقیم ہے۔ (Riaz is resident with his friend in the city of Riyadh.) →(Riaz, ریاض) = <Person> →(Riyadh, ریاض) = <Location>
Inflectional and Agglutinative Morphology	جان نثار وطن کے لیے قربان ہو گئے۔ (The self-sacrificing ones gave their lives for the country.) → (جان نثار) is an adjective meaning "devoted/self-sacrificing", not a named entity. Even though "Nisar" alone can be a name, its use in a compound changes its role. مختار ایک خودمختار ریاست کے قانونی مشیر ہیں۔ (Mukhtar is the legal advisor to a sovereign state.) → "مختار" may refer to a person's name, but when combined with the prefix "خود", it becomes "خودمختار" ("autonomous/sovereign"), no longer a named entity but a descriptive adjective.
Free Word Order	(I opened a new restaurant in Chitral.) میں نے چترال میں ایک نیا ریسٹورنٹ کھولا۔ ایک نیا ریستوران میں نے چترال میں کھولا۔ میں نے ایک نیا ریستوران چترال میں کھولا۔ Flexible syntax in Urdu allows multiple word orders, making entity recognition harder for sequence-based models.
Spelling Variations and Orthographic Ambiguity	A single named entity may appear in multiple surface forms, for example, "سماد", "صمد" and "سمد" all represent the name "Samad" Compound expressions like "کے ساتھ" and "کیساتھ" (both meaning "with") may appear with or without spaces, affecting tokenization and degrading model accuracy.

Figure 5: Key challenges in Urdu Named Entity Recognition, including linguistic ambiguity, morphological complexity, and lack of orthographic cues.