Journal of Computer Science

Research Article

UzNER: A Human-Reviewed Benchmark for Uzbek Named Entity Recognition With Gazetteer-Augmented Transformer Models

Bobur Saidov[1,2], Vladimir Barakhnin[1,3], Zarnigor Fayzullaeva[4], Umid Ibragimov[1] and Ulugbek Tursunov[1]

[1]*Faculty of Mechanics and Mathematics, Novosibirsk State University, Novosibirsk, Russia*
[2]*Faculty of Computer Engineering, Urgench State University, Urgench, Uzbekistan*
[3]*Federal Research Center for Information and Computational Technologies, Novosibirsk, Russia*
[4]*Faculty of Software Engineering, University of Information Technologies, Tashkent, Uzbekistan*

Article history
Received: 21-03-2026
Revised: 30-05-2026
Accepted: 04-06-2026

Corresponding Author:
Bobur Saidov
Faculty of Mechanics and Mathematics, Novosibirsk State University, Novosibirsk, Russia
Email: saidovboburbek9629@gmail.com

Abstract: UzNER-100K is a large-scale human-reviewed benchmark for Uzbek named entity recognition with 100,000 training sentences, 18 fine-grained entity types and 200,083 entity mentions across 114,269 sentences in total. The corpus was constructed through an LLM-assisted, expert-reviewed annotation pipeline that achieved strong reliability on the main audit subset while substantially reducing corpus-construction effort. The benchmark includes a standard test split, a gold-audited subset and a hard subset designed to stress long, ambiguous and structurally complex cases. We evaluate 10 Uzbek NER systems spanning recurrent, monolingual Uzbek, multilingual transformer and hybrid architectures. The best model, XLM-R + Gazetteer + CRF, reaches 91.03 Micro-F1 on the standard test set, 89.67 on the gold-audited subset and 83.21 on the hard subset. Quality control included a dedicated inter-annotator agreement audit, achieving 91.3% span-level agreement, 93.7% entity-type agreement, and a Cohen's Kappa of 0.914. In addition, a qualitative native-speaker assessment confirmed the linguistic naturalness of the model outputs while highlighting remaining challenges in legal, administrative, and event-related expressions.

Keywords: Uzbek NER, Low-Resource NLP, Benchmark Dataset, Multilingual Transformers, Gazetteer-Enhanced Decoding

Introduction

Named Entity Recognition (NER) is a core sequence-labeling problem in natural language processing because it underpins information extraction, retrieval, question answering, knowledge-based construction and many downstream analytics workflows (Li et al., 2022; Lample et al., 2016). Although modern pretrained encoders have greatly improved NER quality in major languages, performance remains substantially more fragile in low-resource settings where annotation scarcity, domain shift and limited coverage of rare entity forms constrain both training and evaluation (Devlin et al., 2019; Conneau et al., 2020; Li et al., 2022).

Uzbek is an important test case in this context. As a morphologically rich Turkic language, it exhibits productive affixation, orthographic variation, mixed-script and transliteration noise and comparatively limited public NLP infrastructure (Abdurakhmonova et al., 2022; Yusufu et al., 2023). These properties make NER challenging not only because entity boundaries are difficult to identify, but also because models trained on narrow datasets often fail to generalize to out-of-domain names, administrative phrases and semi-structured identifiers (Yusufu et al., 2023; Malmasi et al., 2022; Fetahu et al., 2023).

Recent multilingual NER studies have shown that a single held-out split is often not enough to understand robustness (Malmasi et al., 2022; Fetahu et al., 2023). Strong leaderboard scores can coexist with substantial degradation on longer sequences, ambiguous entity types, rarer forms and cross-domain text (Fetahu et al., 2023; Huang et al., 2025). This is especially relevant for Uzbek, where available benchmarks have historically been smaller and more homogeneous than those available for high-resource languages.

This benchmark-oriented study complements recent Uzbek NLP work on hybrid entity- and sentiment-aware modeling, where lexical, deep-learning and entity-based components have already shown strong potential for structured text analysis (Saidov et al., 2026b). In addition to the standard train/dev/test split, the benchmark includes a gold-audited subset and a deliberately hard subset designed to expose robustness gaps. The study also evaluates whether gains come primarily from model scale, from structured decoding or from external lexical knowledge (Ma and Hovy, 2016; Fetahu et al., 2021).

In this context, the present study is positioned not merely as a dataset release, but as a benchmark-oriented investigation of Uzbek NER under realistic evaluation conditions. Our goal is not only to provide a large annotated corpus, but also to examine which modeling choices remain effective when performance is tested beyond a single standard split. To this end, we combine large-scale benchmark construction with comparative modeling, ablation analysis, reviewed synthetic augmentation, and stricter robustness checks across gold-audited, hard, and cross-domain subsets. This framing is important because progress in Uzbek NER should be measured not only by peak in-domain scores, but also by stability, transferability, and resistance to more challenging text conditions.

Beyond benchmark construction and model comparison, this study also investigates several linguistic factors that affect Uzbek NER performance, including suffixation, entity-boundary ambiguity, legal-document references, and event-related expressions. These analyses provide additional insight into the challenges of named entity recognition in morphologically rich low-resource languages.

Related Work

For Uzbek and other Turkic languages, public NER resources are improving but remain comparatively small and heterogeneous in annotation policy, domain coverage and evaluation rigor (Yusufu et al., 2023; Saidov et al., 2025; 2026a). Existing studies demonstrate that transformer encoders are effective, but they also show that low-resource models remain sensitive to domain mismatch and previously unseen surface forms (Devlin et al., 2019; Conneau et al., 2020; Yusufu et al., 2023).

A complementary line of work emphasizes external lexical knowledge. Gazetteers, lexicons and knowledge-aware features can improve named entity recognition when entity spellings are stable enough to provide useful priors, yet not so stable that dictionary matching alone solves the problem (Fetahu et al., 2021). In morphologically rich settings, such resources are most effective when integrated into contextual models rather than applied as a stand-alone rule layer (Fetahu et al., 2021; Li et al., 2022).

Transformer encoders provide the strongest general baseline family for these settings (Devlin et al., 2019; Conneau et al., 2020; He et al., 2023). However, public multilingual models are not trained specifically for Uzbek legal, administrative or news-like phraseology (Conneau et al., 2020; Yusufu et al., 2023). This motivates a hybrid design in which contextual encoding is complemented by structured decoding and lightweight gazetteer features (Ma and Hovy, 2016; Fetahu et al., 2021).

LLM- and Prompt-Based NER

Recent work has also explored large language models and prompt-based methods for named entity recognition. These approaches are attractive because they can perform extraction with limited task-specific training data and can sometimes generalize across domains through instruction following. However, prompt-based NER also introduces several practical challenges for benchmark construction, including prompt sensitivity, output-format instability, higher inference cost, difficulty of exact span alignment, and reduced reproducibility across model versions and decoding settings.

For this reason, the present study focuses on supervised sequence labeling models evaluated under a fixed and reproducible protocol. The proposed XLM-R + Gazetteer + CRF architecture is not intended to replace LLM-based extraction, but to provide a controlled benchmark baseline for Uzbek NER where token-level BIOES labels, exact span matching, multi-seed evaluation, and cross-domain robustness can be measured consistently. A direct comparison with instruction-tuned LLMs and prompt-based NER systems is therefore left for future work.

Materials and Methods

Benchmark Construction and Quality Control

UzNER-100K was designed as a benchmark rather than merely as a conventional train/dev/test release. The core resource contains 114,269 sentences, including 100,000 for training, 2,000 for development, 2,000 for standard testing, 10,000 for gold-audited evaluation, and 269 for a dedicated hard subset. Split integrity was treated as a first-class constraint, and exact duplicate rates were reduced to zero across the reported benchmark splits.

The corpus was assembled through a multi-stage workflow that combines real text collection, LLM-assisted pre-annotation, and human review. Real text was collected from Uzbek public-domain and news-like sources, whereas synthetic sentences were introduced to improve coverage of rare classes, morphologically variable forms, and document-style expressions that remain sparse in naturally occurring text. Every synthetic sample included in the final release was manually reviewed before training.

Synthetic Data Generation and Human Validation

Synthetic data generation was used as a controlled augmentation strategy rather than as a substitute for real Uzbek text. The main purpose of the synthetic component was to increase the coverage of rare entity classes, long multi-token entities, document-style expressions, administrative phrases, and morphologically variable surface forms that are underrepresented in naturally collected data.

The generation process followed a template-and-slot procedure. First, a set of sentence patterns was prepared for different entity contexts, including person mentions, organizations, locations, legal-document references, dates, monetary expressions, events, products, and administrative titles. Second, the slots in these patterns were filled with curated gazetteer entries and manually prepared lexical variants. Third, the resulting sentences were normalized for Uzbek orthography, script consistency, punctuation, and BIOES-compatible entity boundaries. This procedure made it possible to control the intended entity type and span while still allowing lexical and morphological variation.

LLM assistance was used only at the pre-annotation and sentence-generation support stage. Raw generated outputs were not accepted automatically. Each synthetic sentence included in the final benchmark was checked by human reviewers before being used for training or evaluation. During validation, reviewers checked four main criteria:

(i) Whether the sentence was linguistically acceptable in Uzbek
(ii) Whether the entity span boundaries were correct
(iii) Whether the assigned entity type matched the annotation guidelines
(iv) Whether the sentence added useful coverage rather than duplicating already frequent patterns

This design reduces, but does not completely eliminate, the risk of synthetic-data bias. In particular, template-based generation can overrepresent regular sentence structures and may produce cleaner examples than naturally occurring social or legal text. For this reason, synthetic data were used only as a reviewed augmentation layer, while the benchmark also includes real-text evaluation, a gold-audited subset, a hard subset, and cross-domain evaluation. The performance differences across these subsets are therefore important for distinguishing general model robustness from improvements caused by controlled augmentation.

Quality control was incorporated at several stages of benchmark construction. On the audited pilot subset, span agreement reached 91.3% and label agreement 93.7%, indicating that the annotation guidelines were operationally consistent. In addition, the released benchmark was checked for format validity, split leakage, and label consistency prior to publication.

Table 1 summarizes the resulting split structure. The gold-audited subset contains 10,000 manually verified sentences, 140,116 tokens, and 24,832 entity mentions. Its average sentence length is 14.01 tokens, which indicates that it is not simply a collection of unusually long or artificially difficult cases. Instead, it functions as a stricter quality-controlled evaluation regime that complements the standard test split.

The hard subset was designed as a diagnostic stress-test set rather than as a statistically broad standalone benchmark. Its purpose is to expose model behavior on cases that are intentionally more difficult than the standard test distribution. These cases include long sentences, nested or adjacent entity mentions, ambiguous LOC/GPE and ORG/FAC distinctions, administrative expressions, document references, and informal or non-canonical surface forms. Therefore, the hard subset should be interpreted as a qualitative and diagnostic robustness probe. Its relatively small size is a limitation, but it is useful for identifying systematic weaknesses that may be hidden by aggregate performance on the larger standard test set. Future versions of the benchmark will expand this subset to support more fine-grained statistical testing.

As shown in Table 1, the split design supports multiple evaluation regimes rather than a single held-out score. In addition to the conventional test split, the benchmark contains a manually verified gold-audited subset and a hard subset composed of longer and more ambiguous examples. This design reduces the risk of overinterpreting a single benchmark value and makes model comparison more informative.

Selection of the Gold-Audited Subset

The 10,000-sentence gold-audited subset was constructed to provide a stricter and manually verified evaluation regime in addition to the standard development and test splits. Unlike the hard subset, which was intentionally designed to contain difficult and ambiguous cases, the gold-audited subset was not selected through model-uncertainty sampling. Instead, it was created using stratified random sampling from the broader benchmark pool.

Table 1: Benchmark splits and evaluation regimes

Split	Sentences	Tokens	Entity Mentions	Avg. Tokens/Sent.	Role
Train	100,000	939,404	159,442	9.39	Main training split
Dev	2,000	50,031	8,245	25.02	Model selection
Test	2,000	41,611	6,755	20.81	Standard benchmark
Gold-audited	10,000	140,116	24,832	14.01	Stricter verified evaluation
Hard	269	13,264	809	49.31	Stress test for difficult cases

The stratification procedure was designed to preserve the main distributional properties of the full dataset. In particular, the sampling process considered domain coverage, entity-type distribution, sentence length variation, and the presence of both frequent and lower-frequency entity categories. This ensured that common classes such as PER, LOC, DATE, and ORG remained well represented, while domain-specific and less frequent classes such as LAW, DOC, EVENT, MONEY, QUANTITY, and FAC were also included in sufficient numbers for diagnostic evaluation.

After sampling, the selected 10,000 sentences were manually audited by human reviewers. The audit checked entity boundaries, entity types, BIOES label consistency, sentence-level validity, and possible annotation conflicts. Cases involving morphologically marked entity forms, long administrative names, legal-document references, and semantically adjacent classes such as LOC/GPE, ORG/FAC, and LAW/DOC received additional attention during the review.

The purpose of this subset was therefore different from that of the hard subset. The gold-audited subset was intended to serve as a cleaner and more reliable evaluation set that reflects the general benchmark distribution, whereas the hard subset was intended to stress-test model behavior on long, ambiguous, and structurally difficult examples. This design makes it possible to distinguish ordinary benchmark performance from performance under stricter manual verification and targeted difficulty conditions.

Beyond split design, the benchmark is also defined by the breadth of its entity schema and the internal distribution of labeled mentions. The 18-class inventory covers standard named entities and several categories that are especially relevant for Uzbek public, legal, and administrative text, including PER, LOC, ORG, GPE, FAC, LAW, DOC, and EVENT. This broader schema makes the benchmark more realistic than earlier small-scale resources that focus only on the most common entity types.

To characterize the benchmark more concretely, Table 2 reports the entity-type frequencies and average span lengths in the training split. These statistics help clarify both the relative prominence of major categories and the structural diversity of the annotation schema.

As shown in Table 2, LOC, PER, DATE, and ORG are the most frequent categories in the training set, while classes such as LAW, MONEY, EVENT, and DOC are less dominant but still sufficiently represented for controlled evaluation. The average span statistics further indicate that the benchmark contains both compact entities and structurally longer categories, making sequence labeling more realistic and less biased toward single-token mentions. Overall, this distribution supports both standard benchmark comparison and more fine-grained robustness analysis across frequent and less frequent entity types.

Inter-Annotator Agreement and Adjudication

To quantify the reliability of the human validation stage, we conducted an inter-annotator agreement audit on a manually reviewed subset of the benchmark. The audit was designed to evaluate whether the annotation guidelines were sufficiently clear and whether the human-reviewed synthetic samples could be used as reliable training and evaluation material. Two native Uzbek-speaking annotators independently annotated the same validation subset using the BIOES-based annotation scheme adopted for the final corpus.

Agreement was measured at several complementary levels because named entity recognition involves both boundary detection and semantic classification. At the token level, we computed observed agreement over BIOES labels and Cohen's Kappa in order to account for chance agreement. At the span level, we measured whether both annotators selected the same entity boundaries. At the entity-type level, we measured whether the same semantic class was assigned to the corresponding entity mention. Finally, at the entity level, we used exact span-and-type matching, where an entity was counted as agreed only if both the boundary and the entity type were identical.

The overall results of the reliability audit are presented in Table 3. The observed token-level agreement reached 96.18%, while Cohen's Kappa was 0.914, indicating a high level of annotation consistency. The span-level agreement was 91.30%, and the entity-type agreement reached 93.70%. In addition, the exact span-and-type entity-level F1 was 92.15%. These results show that the annotation protocol was stable and that the reviewed synthetic samples were not accepted automatically, but passed through a controlled human validation procedure.

Table 2: Training-set entity distribution in UzNER-100K

Entity Type	Train Count	Share (%)	Avg. Span
LOC	21,956	13.77	1.63
PER	20,809	13.05	1.42
DATE	18,623	11.68	1.55
ORG	15,029	9.43	2.34
POSITION	9,227	5.79	1.64
PRODUCT	8,011	5.02	2.79
FAC	7,879	4.94	3.08
CARDINAL	7,604	4.77	1.00
GPE	5,304	3.33	1.10
TIME	5,000	3.14	3.00
NORP	5,000	3.14	1.12
LAW	5,000	3.14	2.68
MONEY	5,000	3.14	2.89
EVENT	5,000	3.14	2.19
ORDINAL	5,000	3.14	1.16
DOC	5,000	3.14	1.67
PERCENT	5,000	3.14	1.81
QUANTITY	5,000	3.14	2.11
Total	159,442	100.00	1.87

To better understand which entity categories were most reliable and which categories remained more ambiguous, we also computed entity-level agreement separately for each entity type. The results are shown in Table 4. Agreement was highest for relatively regular and well-defined classes such as PER, LOC, DATE, MONEY, and PERCENT. These classes usually have clearer lexical or structural patterns and are less dependent on domain-specific interpretation.

By contrast, lower agreement was observed for DOC, EVENT, PRODUCT, LAW, FAC, and NORP. This pattern is expected for Uzbek NER because these classes often involve semantically broader or domain-specific expressions. For example, legal and administrative texts frequently contain compact references to laws, decrees, articles, certificates, and institutional documents, which may lead to LAW/DOC ambiguity. Similarly, EVENT mentions such as forums, meetings, seminars, or administrative activities may be difficult to separate from common noun phrases unless the broader context is considered. Facility and organization names also create ORG/FAC boundary ambiguity when institutional names include words such as bino, markaz, universitet, majmua, or boshqarma.

The per-class agreement analysis is also useful for interpreting the later model-level error patterns. The same categories that show lower inter-annotator agreement, especially DOC, EVENT, PRODUCT, LAW, and FAC, also appear among the more difficult classes in the automatic NER evaluation. This suggests that the remaining model errors are not simply caused by insufficient model capacity, but are partly related to genuine linguistic and semantic ambiguity in Uzbek legal, administrative, and public-domain text.

After the independent annotation stage, all disagreement cases were resolved through adjudication. The adjudication procedure focused on boundary ambiguity in morphologically marked Uzbek forms, long multi-token administrative names, document references, and semantically adjacent labels such as LOC/GPE, ORG/FAC, and LAW/DOC. The adjudicated labels were then used in the final released version of the benchmark.

Subword Tokenization and Gazetteer Alignment

Uzbek is a morphologically rich and agglutinative language; therefore, named entities frequently appear with case, possessive, locative, ablative, or derivational suffixes. For example, a location or organization name may occur in forms such as *Toshkentda*, *Andijonga*, *universitetining*, or *O‘zbekiston Respublikasi Jinoyat kodeksining*. These surface forms create two related challenges for NER: The entity boundary may be obscured by suffixation, and the gazetteer entry may appear only as a base form while the text contains a morphologically extended form.

Table 3: Inter-annotator agreement results for the audited validation subset

Agreement level	Unit of comparison	Metric	Value
Token-level agreement	BIOES token labels	Observed agreement	96.18%
Token-level agreement	BIOES token labels	Cohen’s Kappa	0.914
Span-level agreement	Entity boundaries	Exact boundary agreement	91.30%
Entity-type agreement	Entity labels	Label agreement	93.70%
Entity-level agreement	Exact span + type	Entity-level Precision	92.09%
Entity-level agreement	Exact span + type	Entity-level Recall	92.20%
Entity-level agreement	Exact span + type	Entity-level F1	92.15%

Table 4: Entity-level inter-annotator agreement by entity type

Entity Type	Annotator 1 Count	Annotator 2 Count	Exact Matches	Precision (%)	Recall (%)	F1 (%)
PER	1,215	1,208	1,166	96.52	95.97	96.24
LOC	1,184	1,176	1,120	95.24	94.59	94.91
DATE	842	836	803	96.05	95.37	95.71
ORG	1,032	1,041	958	92.03	92.83	92.43
GPE	621	613	572	93.31	92.11	92.71
POSITION	438	431	393	91.18	89.73	90.45
FAC	356	361	318	88.09	89.33	88.71
LAW	318	326	286	87.73	89.94	88.82
DOC	304	311	269	86.50	88.49	87.48
EVENT	392	401	337	84.04	85.97	84.99
PRODUCT	476	482	409	84.85	85.92	85.38
MONEY	281	278	260	93.53	92.53	93.03
TIME	219	214	199	92.99	90.87	91.92
CARDINAL	345	351	317	90.31	91.88	91.09
ORDINAL	174	171	158	92.40	90.80	91.59
NORP	228	232	205	88.36	89.91	89.13
PERCENT	143	141	132	93.62	92.31	92.96
QUANTITY	263	268	240	89.55	91.25	90.39
OVERALL	8,831	8,841	8,142	92.09	92.20	92.15

To address this issue, gazetteer matching was performed before XLM-R subword tokenization at the whitespace-token and character-span level. This design allowed the system to identify potential entity mentions in the original text before they were split into SentencePiece subword units. The matching procedure used exact surface-form lookup and normalized lookup for transparent suffixal variants when the base entity form could be recovered reliably. For example, if a gazetteer entry matched the base form *Andijon*, a morphologically marked form such as *Andijonga* could still receive a location-related gazetteer feature when the suffix boundary was transparent.

After gazetteer matching, the original tokens were passed to the XLM-R tokenizer. When a word was split into multiple subword units, the corresponding gazetteer feature vector was propagated to all subwords belonging to that word. This prevented the lexical signal from being lost during subword segmentation. The final token representation used by the model was therefore a concatenation of the contextual XLM-R representation and the aligned gazetteer feature vector before CRF decoding.

Model and Evaluation Protocol

This alignment strategy is important for Uzbek because many entity mentions are not observed only in their dictionary forms. However, the method does not fully solve all morphological ambiguity. No complete morphological analyzer was used in the final pipeline, and suffix-heavy or irregularly written forms can still cause missed entities or boundary errors. This limitation is reflected in the later qualitative error analysis, where morphologically marked forms and long administrative expressions remain among the main sources of residual errors.

The resulting inference procedure is summarized in Algorithm 1. The algorithm shows the conceptual flow from contextual encoding to gazetteer feature construction, feature concatenation, and CRF-based BIOES decoding.

The proposed system combines three components: an XLM-R-base encoder for contextual token representations, a gazetteer matching layer that injects lightweight lexical priors and a CRF decoding layer that improves label-sequence consistency (Conneau et al., 2020; Ma and Hovy, 2016; Fetahu et al., 2021). The study compares this hybrid configuration with recurrent, monolingual Uzbek and multilingual transformer baselines, including BiLSTM-CRF, mBERT, UzBERT, BERTbek, XLM-R-base, XLM-R-large, mDeBERTa-v3-base and XLM-R + CRF (Devlin et al., 2019; Conneau et al., 2020; He et al., 2023; Ma and Hovy, 2016).

All transformer systems were trained under a shared protocol with a maximum sequence length of 256 tokens, BIOES tagging, early stopping on development F1 and a common optimizer family. In addition to the standard test split, models were evaluated on the gold-audited subset, the hard subset and targeted robustness analyses such as learning-curve and multi-seed stability experiments.

Algorithm 1. Hybrid XLM-R + Gazetteer + CRF inference

Input:
s = original Uzbek sentence
G = gazetteer resources

Output:
$\hat{y}$ = predicted BIOES label sequence

1: Perform word-level and character-span gazetteer matching on the original sentence:
$gword \leftarrow GazetteerMatch(s, G)$

2: Tokenize the sentence using the XLM-R SentencePiece tokenizer:
$x = (x_1, x_2, ..., x_n)$

3: Align word/span-level gazetteer features with the corresponding XLM-R subword units:
$g_i \leftarrow AlignGazetteerFeatures(x_i, gword)$

4: Obtain contextual subword representations with XLM-R:
$h_1, h_2, ..., h_n \leftarrow XLMR(x_1, x_2, ..., x_n)$

5: for i = 1 to n do

6: Concatenate contextual and gazetteer features:
$z_i \leftarrow [h_i ; g_i]$

7: end for

8: Apply CRF decoding over the BIOES label space:
$\hat{y} \leftarrow CRFDecode(z_1, z_2, ..., z_n)$

9: Return the predicted label sequence and entity spans:
return $\hat{y}$

Computational Setup and Efficiency Analysis

All neural models were implemented using PyTorch and the HuggingFace Transformers framework. Transformer-based models were trained with a maximum sequence length of 256 tokens, BIOES label encoding, AdamW optimization, and early stopping based on development-set F1. To ensure a fair comparison, all models were trained and evaluated using the same train, development, and test splits.

In addition to accuracy and robustness, we also measured the computational cost of the strongest XLM-

DOI: 10.3844/jcssp.2026.1894.1911

R-based configurations. Table 5 reports the training time, GPU configuration, and average inference latency per sentence. The results show that adding a CRF layer and gazetteer features introduces only a moderate computational overhead. The full XLM-R + Gazetteer + CRF model increases inference latency from 18.2 ms/sentence to 21.4 ms/sentence, while improving Micro-F1 on the standard test set from 88.67 to 91.03. This indicates that the proposed hybrid architecture provides a favorable trade-off between accuracy and efficiency.

As shown in Table 5, the proposed hybrid model remains computationally practical. Compared with plain XLM-R-base, the full model requires only 0.3 additional training hours and adds 3.2 ms to the average inference latency per sentence. This overhead is expected because the gazetteer module performs lightweight token-level matching and the CRF layer adds structured decoding over the BIOES label space. However, the additional cost is relatively small compared with the observed performance gain. Therefore, the proposed model is suitable not only as a high-performing benchmark system but also as a practical architecture for Uzbek NER applications where both accuracy and efficiency are important.

Table 5: Computational efficiency of XLM-R-based model variants

Model	Training Time	GPU	Inference Latency, ms/sentence
XLM-R-base	3.5 h	A100 40 GB	18.2
XLM-R + CRF	3.6 h	A100 40 GB	19.1
XLM-R + Gazetteer + CRF	3.8 h	A100 40 GB	21.4

Results and Discussion

Main Benchmark Comparison

Figure 1 visualizes the main benchmark comparison across model families under three evaluation regimes: the standard test split, the gold-audited subset, and the hard subset. The proposed XLM-R + Gazetteer + CRF model is the best overall system, reaching 91.03 Micro-F1 on the standard test set, 89.67 on the gold-audited subset, and 83.21 on the hard subset. Among the plain encoders, mDeBERTa-v3-base and XLM-R-large are the strongest baselines, but both remain below the proposed hybrid model across the primary evaluation settings.

This benchmark pattern is especially informative for Uzbek NER because it suggests that strong contextual encoders alone are not sufficient to fully resolve the task. The advantage of the proposed XLM-R + Gazetteer + CRF system is best interpreted as the result of combining three complementary sources of information: Multilingual contextual representations, lightweight lexical priors from curated gazetteers, and structured sequence decoding (Conneau et al., 2020; Fetahu et al., 2021; Ma and Hovy, 2016). In a morphologically rich and relatively low-resource setting such as Uzbek, this combination is particularly valuable because many entities are context-sensitive, some categories benefit from stable surface-form evidence, and span consistency still matters under fine-grained BIOES labeling. The main benchmark results therefore do not simply identify the best-performing model; they also indicate which design principles are most effective for high-quality Uzbek NER.

Table 6 reports the exact benchmark values for all compared systems and confirms the same overall ranking shown in Fig. 1.

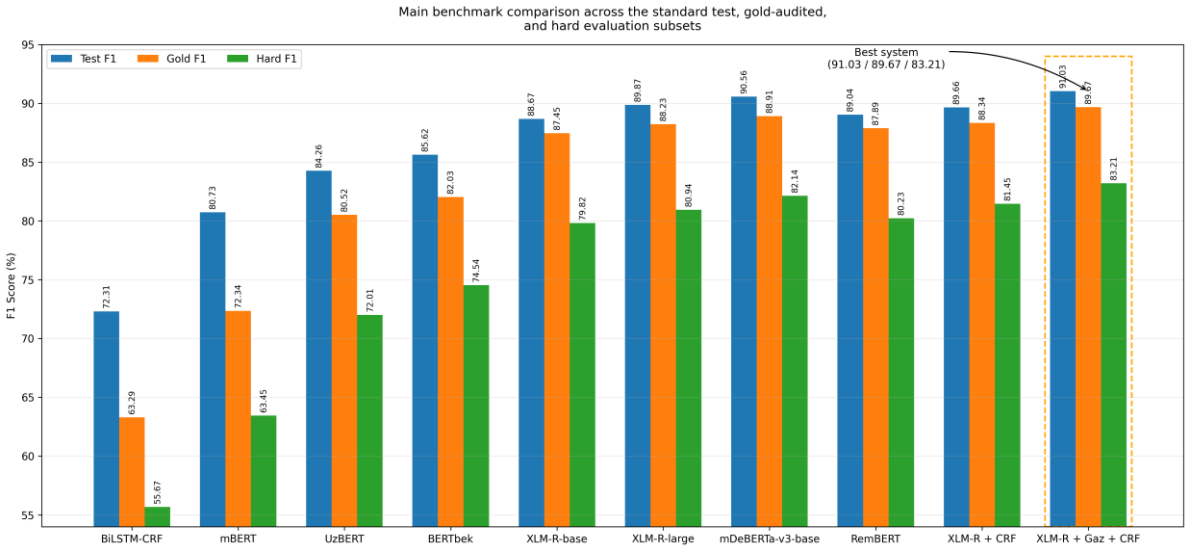

Fig. 1: Main benchmark comparison across the standard test, gold-audited, and hard evaluation subsets. The proposed XLM-R + Gazetteer + CRF model achieves the best overall performance and retains its advantage under stricter evaluation conditions

Table 6: Main benchmark comparison across model families

Model	Family	Test F1	Macro-F1	Gold F1	Hard F1
BiLSTM-CRF	Neural	72.31	68.44	63.29	55.67
mBERT	Transformer	80.73	77.89	72.34	63.45
UzBERT	Transformer	84.25	82.14	80.52	72.01
BERTbek	Transformer	85.63	83.44	82.03	74.54
XLM-R-base	Transformer	88.67	87.34	87.45	79.82
XLM-R-large	Transformer	89.87	88.56	88.23	80.94
mDeBERTa-v3-base	Transformer	90.56	89.28	88.91	82.14
RemBERT	Transformer	89.04	87.78	87.89	80.23
XLM-R + CRF	Hybrid	89.66	88.34	88.34	81.45
XLM-R + Gazetteer + CRF	Hybrid	91.03	89.89	89.67	83.21

Table 7: Contribution of the CRF and gazetteer modules

Variant	CRF	Gazetteer	Precision	Recall	Micro-F1	Macro-F1
XLM-R-base	No	No	89.23	88.12	88.67	87.34
XLM-R + CRF	Yes	No	90.12	89.23	89.66	88.34
XLM-R + Gazetteer	No	Yes	90.34	89.45	89.88	88.56
XLM-R + Gazetteer + CRF	Yes	Yes	91.78	90.34	91.03	89.89

Table 8: Effect of training composition: real versus reviewed synthetic data

Training Setup	Real	Synthetic	Total	Precision	Recall	Micro-F1
Real-only	70K	0	70K	87.23	85.67	86.38
Real + reviewed	70K	10K	80K	88.45	86.78	87.58
Real-heavy mixed	70K	20K	90K	89.67	88.12	88.87
Full mixed (ours)	70K	30K	100K	91.78	90.34	91.03

The stability of the model ranking across standard, gold-audited, and hard evaluation suggests that the proposed improvement is not merely a consequence of favorable test-set composition. Instead, it indicates that lexical priors and structured decoding address complementary weaknesses of multilingual encoders. XLM-R provides broad contextual representations, the gazetteer module improves recognition of recurring and domain-specific entity forms, and the CRF layer reduces invalid or inconsistent BIOES transitions. The remaining performance gap on the hard subset shows that these mechanisms improve robustness but do not fully solve ambiguity, long-span recognition, or domain-specific boundary detection

Ablation and Training Composition

The ablation in Table 7 isolates the two main architectural additions. Starting from plain XLM-R-base, adding a CRF layer alone yields a measurable improvement, while adding the gazetteer alone produces an even larger gain. When both are combined, the score rises to 91.03, showing that structured decoding and lexical priors are complementary rather than redundant.

A second design question is whether the gain comes from architecture alone or also from the mixed-origin corpus. Table 8 shows that training on 70K real sentences only yields substantially lower performance than training on the mixed 100K benchmark. Progressive addition of reviewed synthetic sentences produces consistent improvement, indicating that carefully controlled augmentation can make a meaningful contribution in low-resource Uzbek NER.

Fig. 2 provides a visual summary of the same ablation pattern.

This pattern is methodologically important. The CRF mainly improves sequence consistency and boundary validity under the BIOES scheme, whereas the gazetteer contributes lightweight mention-level lexical evidence for entities with stable surface forms (Ma and Hovy, 2016; Fetahu et al., 2021). Their joint use therefore yields the strongest overall configuration.

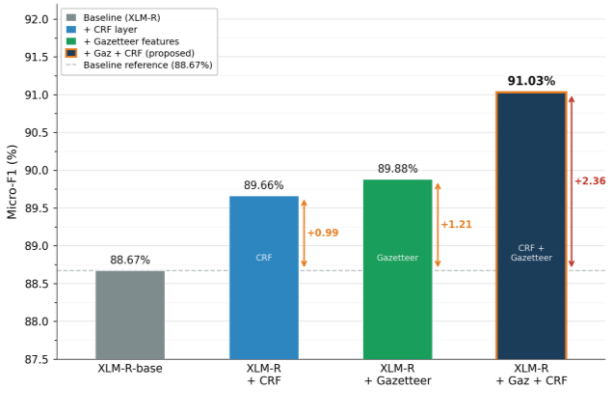

Fig. 2: Ablation study of the CRF and gazetteer components. Both modules improve the XLM-R baseline individually, while their combination yields the strongest overall result

The ablation results are consistent with the linguistic and annotation properties of the benchmark. The CRF decoder primarily helps by enforcing more coherent label transitions and improving span boundary consistency, which is especially relevant under the BIOES tagging scheme. The gazetteer module contributes a different kind of signal: It injects lightweight lexical evidence for entities with relatively stable or recurring surface forms, thereby supporting recall without replacing contextual disambiguation. These gains are therefore complementary rather than redundant. The same logic helps explain the benefit of the mixed training design: Reviewed synthetic sentences do not merely enlarge the corpus, but expand controllable coverage of entity patterns and contextual combinations that would remain sparse in the real-only subset (Huang et al., 2025; Saidov et al., 2026b).

Table 8 and Fig. 3 show that the gain of the final system comes not only from architectural refinement, but also from the mixed-origin training design. Training on 70K real sentences alone yields 86.38 Micro-F1, whereas progressively adding reviewed synthetic data increases performance to 87.58, 88.87, and finally 91.03 in the full 100K mixed configuration.

The trend is important for two reasons. First, it shows that the synthetic portion contributes complementary supervision rather than acting as filler. Second, the gain should be interpreted specifically as the effect of reviewed synthetic augmentation, because the generated sentences were filtered and normalized before inclusion rather than accepted automatically.

Gazetteer Size Sensitivity

In addition to the binary ablation of the gazetteer module, we examined how sensitive the proposed architecture is to the size of the gazetteer resources. This analysis was included because gazetteer coverage may vary substantially across domains, especially for low-resource languages where lexical resources are often incomplete or manually maintained.

For this experiment, we trained the XLM-R + Gazetteer + CRF configuration using progressively larger portions of the gazetteer resources: 25%, 50%, 75%, and 100%. The subsets were constructed through stratified sampling over entity types so that each reduced gazetteer retained partial coverage of major classes such as PER, LOC, ORG, GPE, LAW, DOC, EVENT, and PRODUCT. This setup allowed us to evaluate whether the model depends only on full gazetteer coverage or whether partial lexical resources already provide measurable benefit.

The results are reported in Table 9. The model improves monotonically as gazetteer coverage increases. With 25% gazetteer coverage, the model already improves over the plain XLM-R baseline, indicating that even partial lexical priors are useful. The largest gains appear between 0% and 50% coverage, while the improvement from 75% to 100% is smaller but still positive. This suggests that the gazetteer module is beneficial even when resources are incomplete, but broader coverage remains important for difficult and domain-specific entity types.

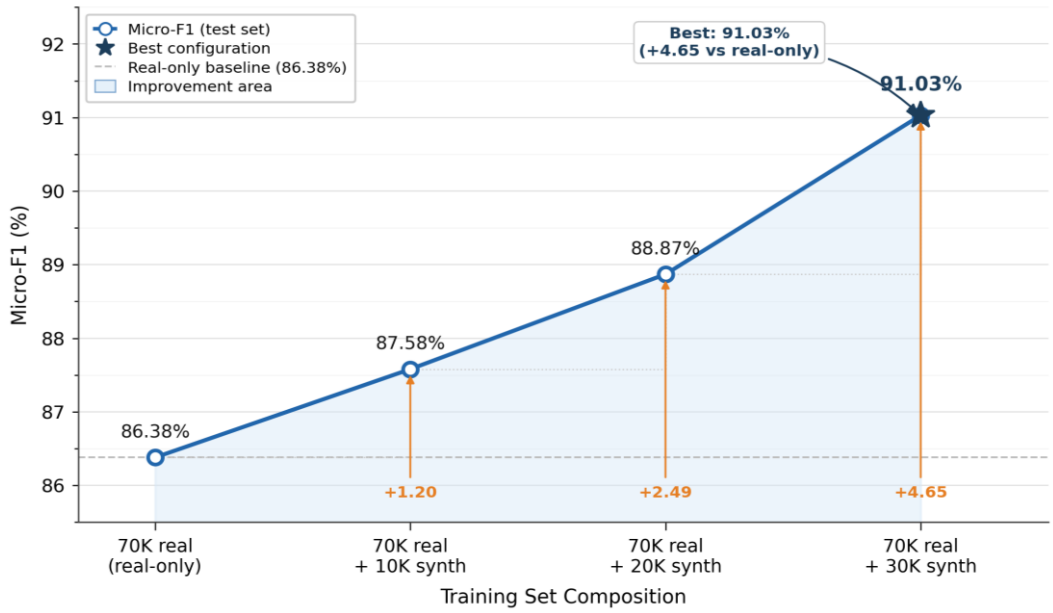

Fig. 3: Effect of progressively adding reviewed synthetic data to the 70K real-sentence training core. Performance improves monotonically, indicating that reviewed synthetic augmentation provides complementary supervision rather than merely increasing corpus size

Table 9: Sensitivity of XLM-R + Gazetteer + CRF to gazetteer resource size

Gazetteer coverage	Precision	Recall	Micro-F1	Difference from XLM-R-base
0% / no gazetteer	89.23	88.12	88.67	0.00
25%	89.94	88.83	89.38	+0.71
50%	90.61	89.42	90.01	+1.34
75%	91.12	89.91	90.51	+1.84
100%	91.78	90.34	91.03	+2.36

As shown in Table 9, gazetteer coverage has a clear but gradually saturating effect. The full gazetteer setting achieves the best result, reaching 91.03 Micro-F1, which corresponds to a +2.36. Point improvement over the plain XLM-R-base model. However, the 50% and 75% settings also provide substantial improvements, showing that the hybrid architecture does not require perfect gazetteer coverage to be useful. This is important for practical Uzbek NER applications because gazetteer resources are likely to remain incomplete in specialized domains such as legal, administrative, product, and event-related text.

The sensitivity analysis also helps explain the remaining error patterns. Classes with more stable lexical realizations, such as PER, LOC, DATE, and ORG, benefit earlier from partial gazetteer coverage, while broader and more domain-dependent classes such as DOC, EVENT, PRODUCT, LAW, and FAC require richer lexical resources and additional contextual disambiguation. Therefore, gazetteer expansion should be viewed as a complementary improvement direction rather than a complete replacement for contextual modeling and structured decoding.

Robustness, Stability and Error Patterns

The benchmark was intentionally designed to reveal robustness gaps. As shown in Table 10, the best system performs best on the standard test split and on cross-domain news, but degrades on the gold-audited, hard, legal and social subsets. This pattern confirms that conventional held-out evaluation overestimates readiness for difficult or domain-shifted Uzbek text.

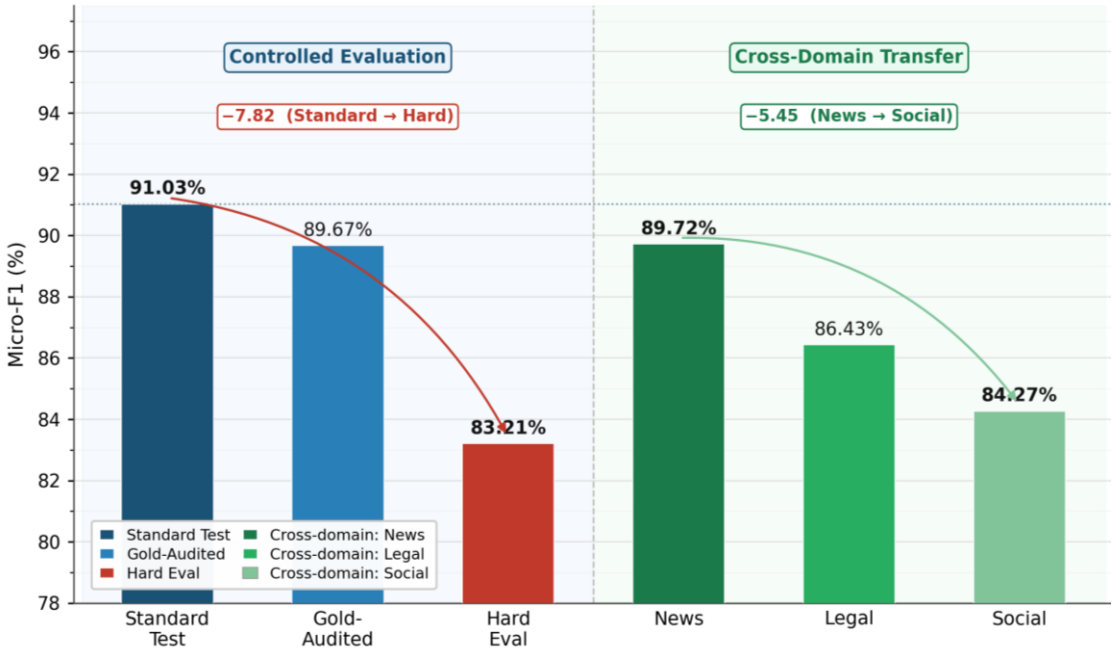

Fig. 4: Performance of the best model across stricter and cross-domain evaluation regimes. The system performs strongest on the standard and news subsets, while hard and social-text evaluation remain substantially more challenging

Table 10: Performance of the best model across stricter and cross-domain evaluation regimes

Evaluation Set	Sentences	Entities	Micro-F1	Avg. Length	Interpretation
Test (standard)	2,000	6,755	91.03	20.81	Main benchmark
Gold-audited	10,000	24,832	89.67	14.01	Stricter manual audit
Hard	269	809	83.21	49.31	Long and ambiguous cases
Cross-domain: News	8,246	22,210	89.72	17.42	Most regular domain
Cross-domain: Legal	490	1,461	86.43	7.84	LAW and DOC heavy
Cross-domain: Social	5,687	14,029	84.27	9.62	Informal short text

Table 10 and Fig. 4 show that the best system remains strong under multiple evaluation regimes, but its performance is not uniform across them. The model achieves 91.03 Micro-F1 on the standard test split, 89.67 on the gold-audited subset, and 83.21 on the hard subset. Under cross-domain evaluation, news is the easiest setting (89.72), followed by legal text (86.43), while social-media text is the most difficult setting (84.27).

This pattern is methodologically important because it shows that a single held-out score overestimates real deployment readiness (Malmasi et al., 2022; Fetahu et al., 2023). The drop on hard and social subsets indicates that long, ambiguous and informal Uzbek text remains substantially more difficult than the canonical benchmark distribution.

The robustness breakdown is informative because the observed performance drop is not random; it reflects systematic differences in text difficulty and annotation conditions. The gold-audited subset is stricter because it reduces tolerance for annotation noise, while the hard subset contains longer, more ambiguous, and less stereotypical entity realizations that are more difficult to recover through local contextual cues alone. Cross-domain variation adds an additional layer of difficulty. Legal text introduces domain-specific terminology and extended nominal structures, whereas social-media text is affected by spelling variation, informal morphology, shortened forms, and reduced orthographic regularity (Fetahu et al., 2023; Yusufu et al., 2023). As a result, the decline outside the standard benchmark should be interpreted as evidence of genuine robustness challenges rather than simple model instability.

To verify that the final ranking does not depend on a single random initialization, Table 11 reports a compact multi-seed comparison for the strongest transformer variants. The best hybrid model varies only modestly across seeds and retains the highest mean F1, which indicates that its advantage is stable rather than accidental.

The stability and data-scaling results further strengthen the interpretation of the benchmark. The multi-seed analysis shows that the proposed configuration remains consistently strong across repeated runs, which reduces the likelihood that the reported best score is an artifact of a favorable random initialization. At the same time, the learning-curve pattern indicates that performance improves in a structured manner as training data increase, rather than saturating prematurely at smaller sample sizes. Taken together, these findings suggest that the model is not only accurate, but also reasonably stable and capable of benefiting from additional supervision. This is important for future Uzbek NER work, because it implies that both further corpus expansion and harder evaluation design are likely to produce meaningful progress rather than merely noisy score fluctuations.

The learning-curve analysis in Table 12 shows that most gains occur in the low-resource and mid-resource regimes, but the benchmark continues to benefit from additional scale up to 100K training sentences. Hard-evaluation performance improves especially strongly with more data, suggesting that rare and structurally complex cases require broad coverage to be learned reliably.

The class-wise analysis in Table 13 shows that the benchmark is not saturated uniformly across entity types. The strongest categories are PER, DATE and LOC, which combine relatively regular lexical realizations with high support in the training set. Lower scores are concentrated in EVENT, PRODUCT and DOC, where class boundaries are semantically broader or surface forms are less standardized.

A compact error audit confirms that the remaining failures are systematic rather than random. The most common category is correct span but wrong class, dominated by semantically adjacent pairs such as LOC↔GPE and ORG↔FAC. Missed entities and boundary truncation are also frequent, especially for long administrative names and low-support legal categories.

Table 11: Multi-seed stability of the strongest transformer variants

Model	Mean F1	Std	Min	Max
XLM-R-base	88.67	0.198	88.43	88.91
XLM-R-large	89.83	0.112	89.67	89.94
mDeBERTa-v3-base	90.52	0.134	90.34	90.67
XLM-R + Gazetteer + CRF	91.01	0.184	90.78	91.23

Table 12: Learning-curve analysis from 5K to 100K training sentences

Train Size	Micro-F1	Macro-F1	Hard F1	Real	Synthetic
5K	70.82	68.23	51.34	3,500	1,500
10K	76.78	74.12	58.67	7,000	3,000
25K	81.92	79.34	65.23	17,500	7,500
50K	85.43	83.67	71.45	35,000	15,000
75K	88.09	86.78	78.23	52,500	22,500
100K	91.03	89.89	83.21	70,000	30,000

Table 13: Per-entity precision, recall and F1 for the best system

Entity	Support	Precision	Recall	F1	Avg. Span
PER	1,342	95.78	94.89	95.33	1.42
DATE	273	94.89	94.12	94.50	1.55
LOC	1,291	93.45	92.78	93.11	1.63
PERCENT	60	93.67	92.34	93.00	1.81
MONEY	314	93.12	91.78	92.44	2.89
GPE	747	92.34	91.23	91.78	1.10
TIME	114	91.56	90.23	90.88	3.00
QUANTITY	60	90.34	89.12	89.72	2.11
ORDINAL	60	89.78	88.56	89.16	1.16
ORG	1,248	89.89	88.45	89.16	2.34
POSITION	214	88.45	87.12	87.78	1.64
CARDINAL	200	88.12	86.78	87.44	1.00
NORP	107	88.67	87.23	87.44	1.12
FAC	157	86.23	84.78	85.49	3.08
LAW	60	86.78	84.56	85.65	2.68
DOC	60	85.67	83.45	84.54	1.67
PRODUCT	206	84.56	82.34	83.43	2.79
EVENT	242	83.12	80.89	81.99	2.19

Table 14: Compact residual error taxonomy for the best system

Error Type	Share (%)	Dominant Labels	Likely Mitigation
Correct span, wrong type	24.9	LOC↔GPE, ORG↔FAC, LAW↔DOC	More contrastive examples; stronger label disambiguation
Missed entity	23.4	EVENT, LAW, DOC, NORP	Tail-class augmentation and oversampling
Boundary too short	18.4	FAC, ORG, PRODUCT, TIME	Morphology-aware tokenization; post-processing
Boundary too long	14.7	ORG, FAC, EVENT	Span regularization and boundary-focused losses
Other / mixed	18.6	Mixed classes	More targeted hard-case review

Qualitative Error Analysis

To complement the quantitative residual error taxonomy in Table 14, we also inspected representative model outputs from the standard, hard, legal, and social-media evaluation subsets. Table 15 presents typical error cases observed for the best-performing XLM-R + Gazetteer + CRF model. The examples illustrate that the remaining errors are not random, but are mostly associated with semantically adjacent entity types, long multi-token spans, document-style expressions, and informal surface forms.

In the table, the gold label denotes the manually corrected annotation, whereas the predicted label denotes the model output. The examples are representative rather than exhaustive and are intended to clarify the main failure modes behind the aggregate error categories reported in Table 14.

As shown in Table 15, many residual errors arise from cases where surface evidence alone is insufficient for reliable disambiguation. LOC/GPE and ORG/FAC errors often involve semantically close categories, while LAW/DOC errors reflect the difficulty of separating legal sources from article or document identifiers. Boundary errors are especially frequent in long administrative and institutional names, where the model may either truncate the full legal form or extend the span to include surrounding descriptive nouns. Social-media examples introduce an additional challenge because informal spelling, suffixation, and shortened expressions may reduce the effectiveness of both contextual encoders and gazetteer matching.

These qualitative examples support the quantitative findings in Table 14. They show that future improvements should focus on contrastive examples for semantically adjacent labels, better handling of morphologically marked entity forms, and additional hard-case annotation for long legal, administrative, and informal expressions.

DOC and EVENT Class Performance in Legal and Administrative Text

The class-wise results show that DOC and EVENT remain among the more difficult categories for the proposed model. As reported in Table 13, the DOC class reaches 84.54 F1, while EVENT reaches 81.99 F1. These scores are lower than those of more regular and highly supported classes such as PER, DATE, LOC, and MONEY. This performance gap is not accidental; it reflects the linguistic and semantic complexity of Uzbek legal and administrative discourse.

The DOC class is difficult because document mentions in Uzbek often appear as compact references embedded inside longer legal or institutional expressions. Examples include references to decrees, orders, certificates, articles, regulations, codes, contracts, and administrative

decisions. In such cases, the boundary between LAW and DOC may be unclear. For instance, in a phrase such as *O‘zbekiston Respublikasi Jinoyat kodeksining 168-moddasi*, the expression *Jinoyat kodeksi* refers to a legal source, whereas *168-moddasi* functions as a document/article-level reference. The model may merge these components into a single LAW span or fail to separate the document reference from the surrounding legal expression.

The EVENT class is also challenging because Uzbek administrative and public-domain texts often express events through common nouns such as *yig‘ilish*, *majlis*, *forum*, *seminar*, *tanlov*, *konferensiya*, or *saylov*. These words may denote either a named event or a generic activity depending on context. For example, *Yoshlar forumi — 2025* should be treated as an EVENT mention, whereas a phrase such as *forumda ishtirok etdi* may function as a common event-related noun phrase rather than a named entity. This semantic ambiguity makes EVENT detection more dependent on context than categories such as PER or DATE.

The lower DOC and EVENT scores are therefore consistent with the residual error analysis in Table 14 and the qualitative examples in Table 15. Many remaining errors involve correct spans with incorrect types, missed event mentions, or boundary errors in long administrative expressions. These results indicate that future improvements for Uzbek NER should include more contrastive examples for LAW/DOC and EVENT/common-noun distinctions, additional legal and administrative annotation, and stronger modeling of document-style and event-style phrase structures.

Table 15: Representative qualitative error examples for the best system

No.	Example sentence	Gold annotation	Model prediction	Error type	Explanation
1	Toshkent viloyati hokimligi yangi qarorni e’lon qildi.	Toshkent viloyati = GPE	Toshkent viloyati = LOC	Correct span, wrong type	The model correctly identified the span but confused an administrative geopolitical unit with a general location. This reflects the common LOC↔GPE ambiguity.
2	Samarqand davlat universiteti binosida ilmiy seminar bo‘lib o‘tdi.	Samarqand davlat universiteti = ORG	Samarqand davlat universiteti binosi = FAC	Boundary too long / wrong type	The model extended the organization span to include “binosi” and shifted the interpretation toward a facility.
3	O‘zbekiston Respublikasi Jinoyat kodeksining 168-moddasi bo‘yicha ish ko‘rib chiqildi.	Jinoyat kodeksi = LAW; 168-moddasi = DOC	Jinoyat kodeksining 168-moddasi = LAW	LAW↔DOC ambiguity	The model merged the legal source and article reference into one LAW span, although the document/article reference should be separated.
4	“Yoshlar forumi — 2025” doirasida yangi loyiha taqdim etildi.	Yoshlar forumi — 2025 = EVENT	Not detected	Missed entity	Event names with punctuation and year markers remain difficult, especially when the phrase resembles an ordinary noun phrase.
5	Navoiy kon-metallurgiya kombinati aksiyadorlik jamiyati yangi mahsulot chiqardi.	Navoiy kon-metallurgiya kombinati aksiyadorlik jamiyati = ORG	Navoiy kon-metallurgiya kombinati = ORG	Boundary too short	The model detected the main organization name but truncated the full legal form of the organization.
6	Kecha Andijonga bordim, lekin net juda yomon ishladi.	Andijonga = LOC	Andijonga = O	Missed entity	Informal or morphologically marked location forms in social-media style text can be missed when the surface form differs from the base gazetteer form.

Table 16: Examples of Uzbek suffixation patterns affecting NER boundary detection

Surface form	Base entity	Suffix/function	Possible NER challenge
Toshkentda	Toshkent	-da, locative	LOC may be missed or suffix may be included inconsistently
Andijonga	Andijon	-ga, directional	Gazetteer match may fail without suffix-aware normalization
Aliyevning	Aliyev	-ning, genitive	PER boundary may include or exclude the suffix inconsistently
universitetining	universitetining	possessive/genitive	ORG boundary may be truncated or overextended
kodeksining	kodeks	-ining, genitive	LAW/DOC boundary may become ambiguous

Morphological Patterns and Boundary Detection in Uzbek NER

A linguistically important source of difficulty in Uzbek NER is the interaction between named entity boundaries and productive suffixation. Uzbek is an agglutinative language, and case, possessive, locative, ablative, genitive, and derivational suffixes are frequently attached directly to named entities. As a result, an entity may appear not only in its base form, but also in morphologically extended forms such as Toshkentda, Andijonga, Aliyevning, universitetining, or kodeksining.

This affects boundary detection in two ways. First, the model must decide whether the suffix belongs to the entity span or should be treated as grammatical material outside the entity. Second, gazetteer-based matching becomes more difficult when the gazetteer contains the base entity form but the text contains a suffixed surface form. For example, Andijon may be present in the gazetteer, while the sentence contains Andijonga. Similarly, Jinoyat kodeksi may occur as Jinoyat kodeksining, where the genitive suffix modifies the surface boundary of the legal reference.

These patterns help explain several residual error types observed in the benchmark. Location and geopolitical names may be missed when they appear with directional or locative suffixes. Organization and facility names may be truncated when the model fails to include the full administrative name, or extended too far when a following common noun is incorrectly absorbed into the entity span. Legal and document references are especially sensitive to suffixation because article numbers, document names, and legal sources often appear in compact morphologically marked forms.

Therefore, suffixation is not only a preprocessing issue, but a core linguistic factor influencing Uzbek NER performance. The proposed gazetteer alignment strategy partially reduces this problem by propagating lexical features from word-level matches to subword units. However, without a full morphological analyzer, some suffix-heavy or irregular forms remain difficult. This is one reason why boundary errors and missed entities continue to appear in the hard, legal, and social-media subsets.

The impact of suffixation on boundary detection can be illustrated through typical Uzbek surface forms. Table 16 presents representative examples in which a named entity appears with a grammatical suffix attached directly to the base form. These examples show why Uzbek NER cannot rely only on dictionary-form matching: The same entity may occur in the text with locative, directional, genitive, possessive, or derivational suffixes. As a result, the model must not only recognize the base entity, but also decide whether the suffix should be included in the entity span or treated as grammatical material outside the named entity.

As shown in Table 16, suffixation affects both lexical matching and span boundary decisions. Location names such as Toshkentda and Andijonga may be missed if the model or gazetteer component only recognizes the base forms Toshkent and Andijon. Person names such as Aliyevning create a different boundary problem because the genitive suffix is grammatically attached to the name, but the named entity itself is usually the base name. Similar ambiguity appears in administrative and legal expressions. Forms such as universitetining and kodeksining may participate in longer ORG, LAW, or DOC mentions, making it difficult to determine the exact entity boundary.

These examples help explain why boundary-related errors remain visible in the hard, legal, and social-media subsets. They also justify the proposed gazetteer-alignment strategy, where gazetteer features are matched at the word or character-span level and then propagated to the corresponding XLM-R subword units. However, as the examples indicate, subword alignment alone does not fully solve the problem of Uzbek suffixation. More robust handling of morphologically marked forms may require suffix-aware normalization or integration of a dedicated Uzbek morphological analyzer in future versions of the system.

Native-Speaker Qualitative Validation

In addition to quantitative evaluation, we conducted a small-scale qualitative validation of model predictions by native Uzbek speakers. The purpose of this assessment was not to replace automatic metrics, but to verify whether the model outputs were linguistically natural and practically interpretable in Uzbek contexts. This step was particularly important because NER errors in Uzbek may involve not only incorrect labels, but also boundary decisions affected by suffixation, administrative phrase structure, and domain-specific terminology.

For this validation, a sample of model predictions was selected from the standard, gold-audited, legal, and social-media subsets. Native Uzbek-speaking reviewers inspected the predicted entity spans and labels according to three criteria:

(i) Whether the predicted entity boundary was natural in Uzbek
(ii) Whether the assigned entity type was semantically appropriate
(iii) Whether the prediction was acceptable in the broader sentence context. Special attention was given to morphologically marked forms, long institutional names, legal-document references, and event-like expressions

The qualitative assessment confirmed that most high-confidence predictions for regular classes such as PER,

LOC, DATE, MONEY, and ORG were linguistically natural and easy to interpret. However, reviewers also identified recurring ambiguity in DOC, EVENT, LAW/DOC, LOC/GPE, and ORG/FAC cases. These findings are consistent with the class-wise results and residual error analysis reported earlier. In particular, native speakers noted that some event expressions may look like ordinary noun phrases unless they contain a specific title, date, or institutional marker, while some legal references require domain knowledge to separate LAW and DOC spans.

This external qualitative check supports the interpretation that the proposed model produces generally reliable Uzbek NER outputs, but also confirms that difficult legal, administrative, and morphologically marked cases require further annotation refinement and more targeted training examples.

The validation criteria and main observations are summarized in Table 17. The table shows that native-speaker assessment largely agrees with the automatic evaluation: Regular entity types are interpreted reliably, while suffix-heavy, legal, administrative, and event-related expressions remain more difficult.

As shown in Table 17, the native-speaker validation did not reveal a general failure of the model, but rather confirmed specific linguistic and domain-related weaknesses. This result strengthens the interpretation that the remaining errors are systematic and concentrated in semantically ambiguous or morphologically complex cases.

Interpretation

The results support three main conclusions. First, multilingual encoders provide a strong baseline for Uzbek NER, but the best results are achieved when contextual modeling is complemented with lexical priors and structured decoding (Conneau et al., 2020; Fetahu et al., 2021; Ma and Hovy, 2016). Second, reviewed synthetic augmentation is useful when it is controlled by a clear schema and human verification (Huang et al., 2025; Saidov et al., 2026a). Third, evaluation beyond a single held-out split is essential for understanding robustness in low-resource settings (Malmasi et al., 2022; Fetahu et al., 2023).

From an application perspective, the benchmark demonstrates that high-quality Uzbek NER is already attainable for news-like and administrative text, but robustness remains lower in legal and social domains. This has immediate implications for deployment in information extraction, search, e-government processing and downstream sentiment or event analysis systems built for Uzbek.

Limitations

Despite the strong overall results, several limitations remain important. First, the training benchmark is mixed in origin, combining 70K real sentences with 30K reviewed synthetic sentences; therefore, part of the observed gain depends on the quality of the generation-and-review pipeline rather than on model architecture alone. Second, cross-domain robustness is still uneven, with social and legal text remaining substantially more difficult than standard and news-like evaluation. Third, the hard subset is highly informative as a diagnostic stress test, but its relatively small size means that it should not be interpreted as a statistically broad standalone benchmark. Finally, the benefit of the gazetteer module depends on lexical coverage and maintenance quality, which may vary across domains and entity types.

Another limitation concerns the possible distributional bias introduced by reviewed synthetic data. Although synthetic examples were manually checked before inclusion, they may still differ from naturally occurring Uzbek text in lexical diversity, discourse structure, spelling noise, and pragmatic variation. This is especially relevant for informal social-media text, where non-standard orthography and shortened forms are common. Therefore, the synthetic component should be interpreted as controlled augmentation rather than as a complete substitute for naturally collected data. The observed performance drop on social and hard subsets confirms that real-world robustness remains an open challenge.

The size of the hard subset is also a limitation. While it is useful for identifying systematic errors in long, ambiguous, and structurally difficult examples, it is not large enough to support all types of fine-grained statistical claims. For this reason, we interpret it as a diagnostic stress-test set and plan to expand it in future versions of UzNER-100K.

Table 17: Summary of native-speaker qualitative validation criteria

Validation criterion	Description	Main observations
Boundary naturalness	Whether the predicted span corresponds to a natural Uzbek entity boundary	High for PER, LOC, DATE; weaker for suffix-heavy and long administrative forms
Entity-type appropriateness	Whether the assigned label matches the semantic role of the mention	Strong for regular classes; ambiguity remains for LAW/DOC, LOC/GPE, ORG/FAC
Contextual acceptability	Whether the prediction is acceptable in the full sentence context	Mostly acceptable in standard/news text; weaker in legal and social-media text
Domain-specific interpretation	Whether legal, administrative, or event expressions require special knowledge	DOC and EVENT decisions often require additional contextual or domain knowledge

A further limitation concerns the scope of the qualitative validation and agreement analysis. Although the benchmark includes an inter-annotator agreement audit and a native-speaker qualitative assessment, both were conducted on selected subsets rather than the entire corpus. Consequently, the reported reliability estimates should be interpreted as strong indicators of annotation quality rather than exhaustive measurements for every benchmark component. Future versions of UzNER-100K will expand both the agreement analysis and external validation procedures to include larger samples and broader domain coverage.

Conclusion

This paper presented UzNER-100K as a large-scale benchmark for Uzbek named entity recognition and evaluated a broad set of baseline and hybrid systems under standard, stricter, and cross-domain conditions. The experiments show that the proposed XLM-R + Gazetteer + CRF model achieves the strongest overall results, while the ablation study confirms that structured decoding and lexical priors provide complementary gains over a strong transformer baseline. In addition, the training-composition analysis demonstrates that reviewed synthetic augmentation contributes meaningful supervision beyond the real-only subset, leading to consistent improvements as the mixed training configuration is expanded.

At the same time, the results show that benchmark-level success should not be interpreted as complete robustness. Performance declines on gold-audited, hard, legal, and especially social-text evaluation, indicating that informal writing, long spans, and domain shift remain open challenges for Uzbek NER. The multi-seed and learning-curve analyses nevertheless suggest that the reported gains are stable and that further progress is realistic with larger, more diverse, and more challenging supervision. Taken together, these findings position UzNER-100K not only as a dataset, but as a practical benchmark for developing and stress-testing high-quality Uzbek NER systems.

The additional analyses provide further evidence that the benchmark is both reliable and linguistically meaningful. The inter-annotator agreement audit demonstrated strong consistency between human reviewers, with high agreement at the token, span, and entity levels. The native-speaker qualitative validation further confirmed that most model predictions are linguistically natural and contextually appropriate in Uzbek. At the same time, the analysis revealed that semantically adjacent classes such as LAW/DOC, LOC/GPE, and ORG/FAC remain challenging for both human annotators and automatic models.

The linguistic analysis showed that Uzbek suffixation and agglutinative morphology continue to influence entity boundary detection, especially in legal, administrative, and social-media text. In addition, the gazetteer sensitivity experiments demonstrated that lexical resources contribute measurable improvements even when only partial coverage is available, while broader gazetteer coverage provides the strongest overall performance. These findings suggest that future progress in Uzbek NER will depend not only on larger transformer models, but also on improved lexical resources, morphology-aware processing, and more diverse evaluation settings.

Future work will focus on expanding cross-domain annotation, increasing the size of the hard subset, improving robustness to informal orthography, and integrating morphology-aware processing for suffix-rich entity forms. We also plan to extend the benchmark with larger legal and administrative corpora, broader gazetteer resources, and more detailed annotation-reliability studies using full kappa-based agreement analysis.

Acknowledgment

The first author acknowledges the support of the El-Yurt Umidi Foundation (Uzbekistan) for academic and research support. This support did not influence the study design, experiments or interpretation of the results. AI assistance was used only for language editing and formatting support; all scientific content, analyses and conclusions were produced and verified by the authors.

Funding Information

This research was partially supported by the state assignment of the Ministry of Science and Higher Education of the Russian Federation for the Federal Research Center for Information and Computational Technologies. The funding number is not applicable/was not provided for this state assignment. The APC was self-funded by the authors.

Author's Contributions

Bobur Saidov: Conceptualization, methodology, dataset design, data curation, software, formal analysis. Experiments, visualization, write original draft, correspondence.

Vladimir Barakhnin: Supervision, methodology review, validation, scientific advising, write review and edited.

Zarnigor Fayzullaeva: Data annotation support, validation; linguistic review, write review and edited.

Umid Ibragimov: Data curation, validation, resource preparation, review of manuscript.

Ulugbek Tursunov: Data preparation, literature support, validation, manuscript review.

Ethics

The released benchmark was prepared for research use. All synthetic sentences were human reviewed before inclusion, and the study did not involve intervention on human subjects. The authors declare that they have no competing interests.

Code Availability

To support transparency and reproducibility, the training and evaluation pipeline used in this study is publicly available at GitHub: https://github.com/Bobur9629/uzbek-ner-model-training.git.

Data Availability Statement

The UzNER-100K benchmark, together with preprocessing scripts, evaluation tools, training configurations, gazetteer-construction utilities, and annotation-support materials, is publicly available on Zenodo: https://doi.org/10.5281/zenodo.18903080.

References

Abdurakhmonova, N., Alisher, I., & Sayfulleyeva, R. (2022). MorphUz: Morphological Analyzer for the Uzbek Language. *2022 7th International Conference on Computer Science and Engineering (UBMK)*, 61–66. https://doi.org/10.1109/ubmk55850.2022.9919579

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Fetahu, B., Chen, Z., Kar, S., Rokhlenko, O., & Malmasi, S. (2023). MultiCoNER v2: a Large Multilingual dataset for Fine-grained and Noisy Named Entity Recognition. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2027–2051. https://doi.org/10.18653/v1/2023.findings-emnlp.134

Fetahu, B., Fang, A., Rokhlenko, O., & Malmasi, S. (2021). Gazetteer-Enhanced Named Entity Recognition for Code-Mixed Web Queries. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1677–1681. https://doi.org/10.1145/3404835.3463102

He, P., Gao, J., & Chen, W. (2023). DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *Proceedings of ICLR 2023*, 1–16. https://doi.org/10.48550/arXiv.2111.09543

Huang, Y., Gao, Y., & Ren, C. (2025). A survey of data augmentation in named entity recognition. *Neurocomputing*, *651*, 130856. https://doi.org/10.1016/j.neucom.2025.130856

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 260–270. https://doi.org/10.18653/v1/n16-1030

Li, J., Sun, A., Han, J., & Li, C. (2022). A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50–70. https://doi.org/10.1109/tkde.2020.2981314

Ma, X., & Hovy, E. (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)* (pp. 1064–1074). https://doi.org/10.18653/v1/p16-1101

Malmasi, S., Fang, A., Fetahu, B., Kar, S., & Rokhlenko, O. (2022). SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER). *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 1412–1437. https://doi.org/10.18653/v1/2022.semeval-1.196

Saidov, B. R., Barakhnin, V. B., Rixsibayev, U. T., Sobirov, O. O., Bekchanov, K. M., & Sharipov, E. J. (2025). Methods of Automatic Selection of Named Entities (NER) in Uzbek Language for Text Tone Analysis. *2025 IEEE 26th International Conference of Young Professionals in Electron Devices and Materials (EDM)*, 1740–1745. https://doi.org/10.1109/edm65517.2025.11096748

Saidov, B., Barakhnin, V., Madirimov, S., Ibragimov, U., Meylikulov, S., Normamatov, S., Bahodirova, F., Matnazarov, J., & Fayzullaeva, Z. (2026a). Dual-Source Synthetic Uzbek Corpora for Sentiment Analysis and NER with Controlled Emoji Signals. *Data*, *11*(2), 28. https://doi.org/10.3390/data11020028

Saidov, B., Barakhnin, V., Saparbaev, R., Narmuratov, Z., Manzura, R., Zilolakhon, R., & Atajanova, A. (2026b). A Hybrid NER-Sentiment Model for Uzbek Texts: Integrating Lexical, Deep Learning, and Entity-Based Approaches. *Big Data and Cognitive Computing*, *10*(3), 92. https://doi.org/10.3390/bdcc10030092

Yusufu, A., Jiang, L., Ainiwaer, A., Teng, C., Yusufu, A., Li, F., & Ji, D. (2023). UZNER: A Benchmark for Named Entity Recognition in Uzbek. *Natural Language Processing and Chinese Computing*, *14302*, 171–183. https://doi.org/10.1007/978-3-031-44693-1_14