informatics

MDPI

Article

Cross-Lingual Transfer of Named Entity Markup with Large Language Models

Vladimir Barakhnin [1,2], Rustam Mussabayev [3], Davlatyor Mengliev [4,*], Alexander Krassovitskiy [3], Alymzhan Toleu [3], Daniil Lyutaev [2], Iskander Akhmetov [3] and Bahodir Ibragimov [5]

[1] Federal Research Center for Information and Computational Technologies, Novosibirsk 630090, Russia
[2] Department of Information Technologies, Novosibirsk State University, 2 Lyapunova Str., Novosibirsk 630090, Russia
[3] Laboratory for Analysis and Modeling of Information Processes, Institute of Information and Computational Technologies, 28, Shevchenko Str., Almaty 050010, Kazakhstan; i.akhmetov@ipic.kz (I.A.)
[4] Department for Scientific Researches, Cyber University, 42, Yangiobod Str., Nurafshon 111500, Uzbekistan
[5] Department of Information Technologies, Urgench State University, 14, Kh. Alimdjan, Urgench 220100, Uzbekistan
* Correspondence: d.mengliyev@csu.uz; Tel.: +998-937-511-616

Abstract

This paper investigates the problem of cross-lingual named entity recognition (NER), which involves automatically identifying entities such as persons, organizations, locations, and other structured elements in text. High-quality NER typically requires manually annotated corpora; however, for many low-resource languages, such data are scarce and costly to produce. The study addresses the following question: can annotated sentences in one language be used to transfer NER markup to their machine-translated counterparts in other languages? To explore this, we propose an approach based on a large language model (LLM) that performs two tasks simultaneously: translating a source sentence and generating BIOES-formatted entity tags for the translated output. To improve robustness and reduce semantic drift, a back-translation step is incorporated to verify meaning preservation by comparing the reconstructed source sentence with the original. The proposed method is compared with two baseline approaches: (1) annotation projection via machine translation and (2) automatic tagging using pre-existing NER tools. Performance is evaluated using standard metrics, including precision, recall, and F1-score. Experimental results demonstrate that the LLM-based approach provides a practical and efficient mechanism for transferring NER annotations across languages. While the method achieves strong and balanced performance, its quality remains influenced by translation accuracy and adherence to annotation constraints. Methodologically, the approach can be considered relatively language-independent, as it relies on general LLM capabilities, a universal tagging scheme, and multilingual semantic representations rather than language-specific model training.

Keywords: cross-lingual transfer; annotation transfer; named entity recognition; multilingual markup; large language models; prompt engineering; BIOES tagging scheme; low-resource languages; back-translation

Academic Editor: Antony Bryant

Received: 11 February 2026
Revised: 16 April 2026
Accepted: 25 April 2026
Published: 7 May 2026

1. Introduction

Today, many text processing tasks begin with understanding what the text is actually about [1,2]. Text often contains important names: people, organizations, cities and countries, dates, documents, and other similar elements [3,4]. The task of automatically identifying

and highlighting such elements is called named entity recognition (NER) [5]. It is useful for information retrieval, news analysis, building knowledge bases, and other practical tasks [6,7]. In addition, it should be emphasized that the NER typically requires labeled data [8,9]. This means that the texts are pre-labeled to identify which words refer to entities and what type they are. Manual labeling is difficult: it takes a long time and requires people with a good knowledge of the language and labeling rules [10]. For low-resource languages, NER progresses more slowly because labeled corpora are limited and manual annotation is expensive and time-consuming.

This study addresses a low-resource scenario in which a BIOES-labeled Uzbek corpus is available, while comparable labeled data for Russian and English are limited. We investigate whether span-level annotations can be transferred from Uzbek to these target languages in an efficient and reliable way.

We adopt the BIOES tagging scheme, which marks the beginning, inside, end, and single-token entities, as well as non-entity tokens.

We propose an LLM-based pipeline that translates a sentence into the target language and generates BIOES tags for the translated tokens. This approach does not require training a separate NER model for each language and can be applied when labeled data are limited.

To control translation faithfulness, we apply back-translation and semantic similarity checks. Although this verification is not perfect, it helps identify where the translation substantially distorts the original meaning.

We compare the proposed LLM-based pipeline with two baselines: (i) annotation projection via machine translation and (ii) automatic tagging of translated text using the Stanza NER toolkit.

We evaluate the transferred annotations using standard NER metrics (precision, recall, and F1) and analyze common error types, including boundary errors, type confusions, and translation-related mismatches.

Overall, the goal of the paper is to assess whether LLMs can accelerate cross-lingual BIOES annotation transfer from Uzbek to Russian and English and how this approach compares to simple baselines.

Contributions:

(1) An LLM pipeline for Uzbek→RU/EN BIOES transfer with prompt constraints + BIOES consistency checks;
(2) An embedding-based semantic analysis (MiniLM + cosine) at sentence level and entity level;
(3) An evaluation on expert reference sets (300 + 300) + comparison with projection and Stanza + error taxonomy.

The objectives of this paper are:

(1) to develop an LLM-based cross-lingual BIOES transfer pipeline for Uzbek →Russian/English;
(2) to compare it with translation-based projection and a ready-made NER baseline;
(3) to evaluate quality on expert-labeled reference sets and analyze typical errors.

The rest of the paper describes the data and methods, then presents the results and discusses what worked well and what needs improvement.

2. Related Works

This section describes the approaches commonly used when NER is needed and when labeled data is insufficient.

2.1. NER and the Problem of Labeled Data

Many modern NER methods work well only when there are many labeled examples [11–13]. If labeling is scarce, quality typically declines. Therefore, for languages without large corpora, researchers often look for ways to obtain labeling faster or transfer it from another language.

2.2. Transferring Labeling via Translation (Projection)

One of the most straightforward methods is to translate sentences into another language and attempt to transfer entity labels to the translation [14,15]. This approach is often called "labeling projection."

However, there are problems with this:

(1) The translation may change word order or word form;
(2) Sometimes an entity is translated not literally but by meaning;
(3) Because of this, it is difficult to accurately determine where the entity is located in the translation and where its boundaries are.

2.3. Ready-Made Tools and Models

Ready-made NLP libraries and pretrained models can also be used for NER [16,17]. This option is attractive because it can produce results quickly; however, performance depends on (i) language coverage, (ii) domain similarity between training data and the target texts, and (iii) the availability of the required entity types.

2.4. Large Language Models (LLM) for Annotation

In recent years, large language models have become popular due to their ability to follow natural-language instructions for many NLP tasks [18,19]. This enables another option: asking the model to directly output token-level annotations, for example, in BIO or BIOES format [20,21].

The advantage of this approach is that it can be applied even when there is no dedicated, well-trained NER model for the target language.

However, LLM outputs may contain formatting errors, missed entities, or incorrect boundaries; therefore, verification and quality control are required.

2.5. How Is the Approach in This Work Different?

The idea of this work is to use an LLM to translate a sentence, immediately generate BIOES tags, and then apply back-translation and semantic similarity checks. We view this as a practical way to obtain cross-lingual span-level annotations faster than fully manual labeling.

More broadly, our work is related to studies of cross-domain adaptation and transfer, where a model or representation learned in one setting is reused in another. Although recent works in areas such as domain generalization, cross-domain segmentation, and cross-domain recommendation explore similar transfer-oriented ideas, their methods are not directly applicable to token-level NER annotation transfer because they operate on different data structures, supervision types, and target outputs. In contrast, our task requires preserving token boundaries and entity labels across languages, which makes translation faithfulness and span consistency central concerns.

The next section will detail the data and steps of the proposed method, followed by a presentation of the results and a comparison with baseline approaches.

3. Materials and Methods

3.1. Source Data

We use an Uzbek corpus of approximately 10,000 sentences annotated at the token level with the BIOES scheme. Each token is assigned a BIOES tag (B/I/E/S/O) together with the corresponding entity type, indicating whether the token belongs to a named entity and its position within the entity span. In addition, it should be noted that Figure 1 reflects input data, whereas Figure 2 demonstrates output data. Table 1 summarizes the datasets used in the study.

Sentence	Word	BIOES-Tag
Sentence 1	O'zbekiston	B-LOC
Sentence 1	Respublikasi	E-LOC
Sentence 1	Prezidentining	S-PER
Sentence 1	2023	O
Sentence 1	yil	O
Sentence 1	25	O
Sentence 1	yanvardagi	O
Sentence 1	Respublika	S-LOC
Sentence 1	ijro	O
Sentence 1	etuvchi	O
Sentence 1	hokimiyat	B-ORG
Sentence 1	organlari	E-ORG

Figure 1. Example of input data: Uzbek sentence tokens with BIOES labels (span-level annotation).

Table 1. Summary of datasets used in the study.

Dataset	Language	Size (Sentences)	Annotation Format	Who Annotated	Role in Experiments	Notes
Source BIOES corpus	Uzbek (source)	≈10,000	span-level BIOES (B/I/E/S/O) + entity type per token	Expert annotation	Input to all pipelines (Uzbek → RU/EN transfer)	Example shown in Figure 1 (input format)
Reference set (gold)	Russian (target)	300	span-level BIOES + entity type per token (used as gold)	Expert annotation	Evaluation (strict span-level and semantic evaluation)	Used for Tables/Figures of results
Reference set (gold)	English (target)	300	span-level BIOES + entity type per token (used as gold)	Expert annotation	Evaluation (strict span-level and semantic evaluation)	Used for Tables/Figures of results

All compared approaches take Uzbek sentences as input and produce span-level BIOES annotations in the target language (Russian or English). The goal is not only to translate the text but also to obtain a target-language annotation that follows the same BIOES structure and can be used for downstream training and evaluation.

For evaluation, we prepared expert-annotated reference sets containing 300 sentences for Russian and 300 sentences for English. Entity types: PER (person), ORG (organization), and LOC (location).

Sentence	Word_original_uz	BIOES-Tag_original_uz	BIOES-Tag_ru	Word_en	BIOES-Tag_en
Sentence 9	Xorazm	B-LOC	O	In	O
Sentence 9	viloyatidagi	E-LOC	S-PER	the	O
Sentence 9	Professorlar	S-PER	O	Khorezm	B-LOC
Sentence 9	yuklamasi	O	B-LOC	region	E-LOC
Sentence 9	o'quv	O	E-LOC	the	O
Sentence 9	va	O	O	workload	O
Sentence 9	ilmiy	O	O	of	O
Sentence 9	faoliyatga	O	O	Professors	S-PER
Sentence 9	o'quv	O	O	is	O
Sentence 9	jarayonlari	O	O	directed	O
Sentence 9	esa	O	O	towards	O
Sentence 9	talabalarda	S-PER	O	educational	O
Sentence 9	mustaqil	O	O	and	O
Sentence 9	ta'lim	O	O	scientific	O
Sentence 9	olish	O	O	activities	O
Sentence 9	ko'nikmasini	O	O	while	O
Sentence 9	rivojlantirishga	O	O	the	O
Sentence 9	yo'naltiriladi	O	O	educational	O
Sentence 9			O	processes	O
Sentence 9			O	are	O
Sentence 9			O	aimed	O
Sentence 9			O	at	O
Sentence 9			O	developing	O
Sentence 9			S-PER	the	O
Sentence 9				students'	S-PER
Sentence 9				skills	O
Sentence 9				for	O
Sentence 9				independent	O
Sentence 9				learning	O
Sentence 10	Qashqadaryo	B-LOC	O	The	O
Sentence 10	viloyatidagi	E-LOC	O	proposal	O
Sentence 10	Oliy	B-ORG	O	regarding	O
Sentence 10	ta'lim	I-ORG	O	the	O
Sentence 10	fan	I-ORG	O	introduction	O
Sentence 10	va	I-ORG	O	of	O

Figure 2. Example of output data: translated tokens in the target language with BIOES labels produced by the method.

3.2. Compared Approaches

This paper compares three methods for obtaining NER annotations in the BIOES format for Russian and English based on Uzbek sentences. All approaches start with Uzbek sentences, but the methods then proceed differently.

It is important to emphasize the difference between the two basic methods:

(1) In the annotation projection approach (Section 3.2.2), we attempted to transfer entities from the original Uzbek annotation to the translation;
(2) In the Stanza baseline approach (Section 3.2.3), entities in the translation are newly discovered by the existing model, meaning the original Uzbek annotation is not directly used.

3.2.1. Proposed LLM Approach

The main approach uses a large language model to obtain BIOES tagging in the target language in a more direct way. In this work, we used ChatGPT (OpenAI, GPT-5.2) as the LLM in the experimental pipeline. The model was used to (i) translate Uzbek sentences into Russian and English and (ii) generate BIOES tag sequences under a fixed prompt template. A query is generated for the model to perform two tasks simultaneously:

(1) translate an Uzbek sentence into Russian or English;
(2) produce BIOES tags for words in the translated sentence.

To ensure the model produces more stable results, simple and clear rules were given in the prompt:

(1) which BIOES tags are allowed (B-, I-, E-, S-, O);
(2) that entity boundaries must be preserved;
(3) that the response must be in a specified format (e.g., "token—tag").

After receiving the response, two checks are performed.

Test 1—BIOES format and logic.

We check that each word has a tag and that the tagging is consistent (e.g., there is no E without a B or I without the beginning of an entity).

Test 2—back-translation and sentence-level semantic similarity.

To reduce the risk of meaning distortion, the resulting Russian/English sentence is translated back into Uzbek using the same LLM (ChatGPT). We then compute sentence embeddings with the multilingual Sentence-Transformer model sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 and measure cosine similarity between the original Uzbek sentence and its back-translation. These similarity scores are used for descriptive analysis of translation faithfulness (reported in Table 2 and Figure 3), but they are not used as a hard filtering threshold in the final evaluation.

Table 2. Semantic similarity statistics for back-translation and direct translation (Russian and English).

Statistic	Back-Translation (RU)	Back-Translation (EN)	Direct Translation (RU)	Direct Translation (EN)
mean	0.887	0.8605	0.9607	0.9267
median	0.9268	0.8972	0.9783	0.9452
min	0.6004	0.5068	0.8574	0.6721
max	0.9751	0.9731	1.0	1.0

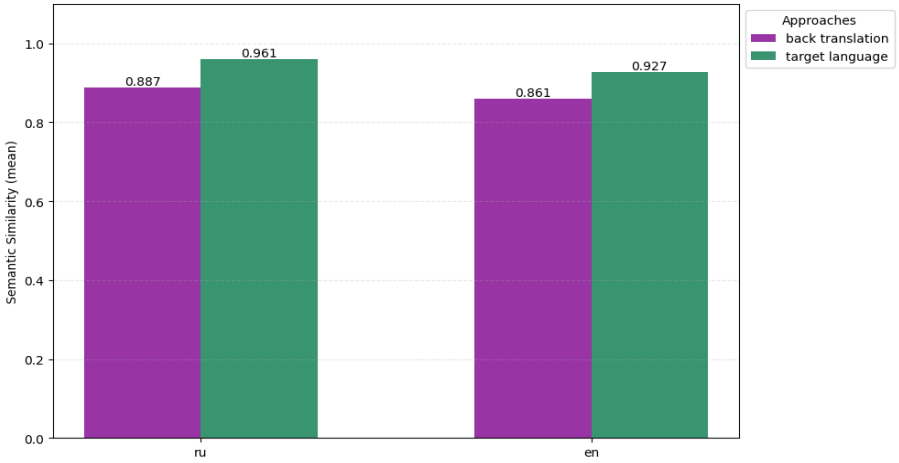

Figure 3. Semantic similarity comparison: original vs. back-translation and original vs. direct target-language translation (Russian and English).

The output of this approach is a sentence in the target language and a BIOES annotation obtained from the LLM after these checks. Figure 4 summarizes the workflow of the proposed approach.

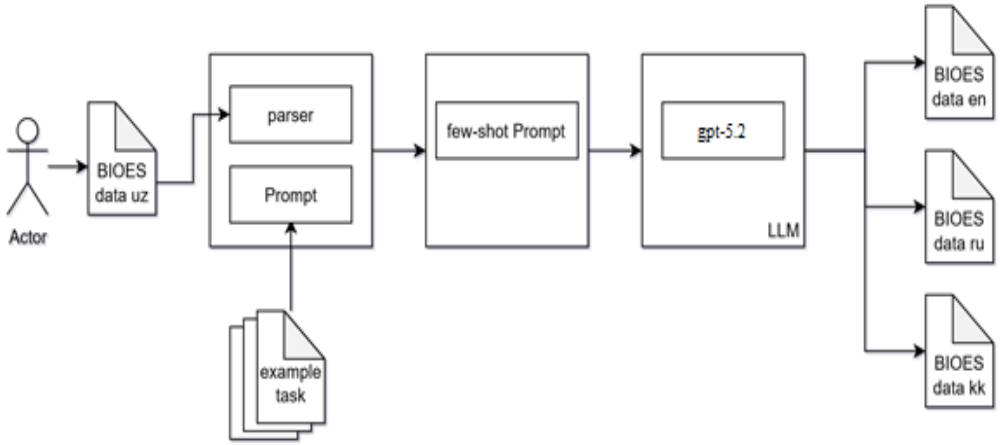

Figure 4. Workflow of the proposed LLM-based pipeline (translation, BIOES tagging, format validation, and back-translation/semantic similarity analysis).

3.2.2. The Approach of Markup Transfer via Translation

The second approach uses the classic idea of markup transfer, where entities from the Uzbek sentence appear in the translation at the same places according to their meaning, so that BIOES tags can be added to the Russian or English text.

The main difficulty here is that during translation:

(1) words may change order;
(2) entities may not be translated literally;
(3) entities sometimes become shorter or longer.

To more easily identify entity boundaries in the translation, an auxiliary technique with tags was used.

Method steps:

(1) Take an Uzbek sentence and highlight each entity in turn with special tags (e.g., <ent>...</ent>).
(2) Translate this sentence into the target language.
(3) Based on the position of the tags in the translation, we determine which text fragment corresponds to a given entity.

Then, we take the translation of a regular sentence (without tags) as the final text and transfer the identified entity boundaries into it.

After this, we add BIOES tags to the words in the final translation.

It is important to note that in this approach, we do not aim to find new entities, but rather to transfer those already tagged in the Uzbek text. In addition, for a fair comparison, the same translation system was used in all pipelines (ChatGPT 5.2).

3.2.3. Approach with a Pre-Existing NER Model (Stanza Approach)

The third approach is used as a baseline for comparison. Here, the original Uzbek markup is not transferred, and entities are not mapped one-to-one. Instead, a fairly simple scheme is used: first, the text in the target language is obtained, and then it is annotated with a ready-made NER model.

Method steps:

(1) The Uzbek sentence is translated into Russian or English.
(2) The ready-made NER model from the Stanza library is run for the corresponding language.
(3) The model outputs a list of entities and their types found in the translation.
(4) The result is converted to BIOES format, meaning B/I/E/S/O labels are assigned to the words in the sentence.

The main idea of this approach is that Stanza finds entities anew, so the result may differ from the original Uzbek markup. This makes the method useful as a baseline: it shows what can be achieved out of the box without transferring the markup.

3.3. Evaluation Methodology

All methods were evaluated on expert-labeled reference sets consisting of 300 sentences in Russian and 300 sentences in English. Evaluation was performed at the entity span level. We report micro-averaged Precision, Recall, and F1-score. Under strict matching, a predicted entity is counted as a True Positive (TP) only if its span boundaries and entity type exactly match the reference. A predicted entity that does not match any reference entity is counted as a False Positive (FP), and any reference entity not matched by predictions is counted as a False Negative (FN).

In addition to strict span-level evaluation, we report two complementary analyses to capture translation effects: (i) sentence-level semantic similarity for meaning preservation, and (ii) entity-level semantic matching to account for surface-form variation.

Sentence-level semantic similarity (meaning preservation)

For the LLM-based pipeline, we assess meaning preservation using back-translation. Each generated target-language sentence (Russian/English) is translated back into Uzbek using the same LLM (ChatGPT 5.2). We compute sentence embeddings with the multilingual Sentence-Transformer model sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2, which maps different languages into a shared embedding space, and measure cosine similarity between the original Uzbek sentence and its back-translated version. We also compute cosine similarity between the original Uzbek sentence and the direct Russian/English translation. Descriptive statistics and score distributions are reported in Table 2 and Figure 4. These similarity scores are used for analysis only and are not applied as a hard filtering criterion in the final evaluation.

Entity-level semantic matching (semantic equivalence in evaluation)

To reduce penalties for translation-induced surface-form variation (e.g., abbreviations vs. full forms or official vs. shortened names), we also compute entity-level scores allowing semantic matching. Entity embeddings are computed using the same model (paraphrase-multilingual-MiniLM-L12-v2), and cosine similarity is calculated between predicted and reference entity strings.

Semantic matching is applied under the following constraints:

1. Same entity type only (e.g., LOC↔LOC, ORG↔ORG);
2. Matching is performed within the same sentence;
3. A predicted entity is counted as a semantic TP if the maximum cosine similarity with any reference entity of the same type in that sentence is $\geq$T_ent = 0.85;
4. To avoid multiple predictions matching the same reference entity, one-to-one matching is enforced by greedily assigning the highest-scoring pairs.

We set T_ent = 0.85 based on commonly used similarity thresholds for multilingual sentence embeddings and confirmed in a small pilot check that it provides a reasonable precision–recall trade-off. All reported metrics are computed at the entity-span level (not

token accuracy). This setup ensures that both boundary detection and correct entity typing are properly evaluated.

3.4. Illustrative Example

To illustrate typical cross-lingual transfer challenges, we provide parallel examples in Uzbek, Russian, and English. For clarity, we include a short gloss (approximate meaning) and show BIOES tags for the main entity span.

Example 1.

- Uzbek: "Men Axborot va hisoblash texnologiyalari institutiga bordim."
- Gloss: "I went to the Institute of Information and Computational Technologies."
- Russian: "Я пошёл в Институт информационных и вычислительных технологий."
- English: "I went to the Institute of Information and Computational Technologies."

Target BIOES annotation (ORG)—correct span:

- Uzbek tokens: Axborot B-ORG, va I-ORG, hisoblash I-ORG, texnologiyalari I-ORG, institutiga E-ORG;
- Russian tokens: Институт B-ORG, информационных I-ORG, и I-ORG, вычислительных I-ORG, технологий E-ORG
- English tokens: Institute B-ORG, of I-ORG, Information I-ORG, and I-ORG, Computational I-ORG, Technologies E-ORG

Typical error: The model labels only the head token (e.g., "Institute/Институт") as ORG, missing the full multi-word span. This produces a strict span mismatch even when the entity type is correct.

Example 2. Abbreviation vs. full form.

Uzbek: "Bugun AQSH Prezidenti bayonot berdi."

Russian: "Сегодня президент США сделал заявление."

English: "Today the U.S. President issued a statement."

Reference entity: AQSH/США/U.S. (LOC)

Typical challenge: Surface forms differ across languages (abbreviation, punctuation, and transliteration). Under strict evaluation, such differences may reduce scores if boundary/type alignment is imperfect. Under semantic evaluation, the case can be matched if the entity type is the same and the cosine similarity exceeds T_ent.

Example 3. Variation in official vs. short name.

Uzbek: "Uchrashuv O'zbekiston Respublikasi Prezidenti bilan bo'lib o'tdi."

Russian: "Встреча состоялась с Президентом Республики Узбекистан."

English: "The meeting was held with the President of Uzbekistan."

Reference entity: O'zbekiston Respublikasi (LOC)

Typical challenge: The English translation may use a shortened form ("Uzbekistan") instead of the full official form ("Republic of Uzbekistan"). In strict span-level evaluation, this leads to a mismatch, while semantic evaluation can treat it as correct when the entity type matches and the cosine similarity is $\geq$T_ent.

4. Results

This section discusses the results of a comparison of three approaches: the LLM approach, annotation projection, and a ready-made model (Stanza). The evaluation is conducted on a reference set of 300 sentences with expert BIOES annotation for each target language (Russian and English). Below, we examine (1) the preservation of meaning separately after translation and (2) the quality of NER annotation using standard metrics.

4.1. Semantic Similarity (Preservation of Meaning and Back-Translation)

The LLM-based approach additionally evaluates how well the meaning of a sentence is preserved during translation. This is particularly important because the LLM pipeline performs translation and BIOES annotation simultaneously; therefore, translation errors may directly affect entity labeling.

Two semantic consistency comparisons are performed:

- Original vs. Back-Translation—comparison between the original Uzbek sentence and its back-translation (after translation into Russian/English and back into Uzbek);
- Original vs. Target Translation—comparison between the original Uzbek sentence and its translation into Russian or English.

In addition, Figure 4 illustrates the distribution of similarity values, and Table 2 provides descriptive statistics (mean, median, minimum, and maximum).

Overall, the results show that semantic similarity values are generally high across most sentences. This indicates that the LLM approach preserves overall sentence meaning in the majority of cases. Lower similarity values are observed primarily in structurally complex sentences or in cases involving ambiguous constructions. We embed Uzbek and Russian/English sentences in the same multilingual embedding space (MiniLM) and compute cosine similarity.

For the annotation projection and Stanza approaches, no back-translation quality control was applied. Although machine translation is used as a technical step in these pipelines, it is not explicitly evaluated through semantic consistency checks.

4.2. Quality of BIOES Tags Without Regard to Semantics

Next, we evaluate how accurately each method reproduces the original BIOES annotation under strict span-level criteria. An entity is considered correct only if:

- the entity boundaries (span) exactly match the reference, and
- the entity type (label) is identical.

The results are presented in Table 3 and visualized in Figures 5–7, which report Precision, Recall, F1-score, and TP/FP/FN distributions. It is important to note that the strict results in Table 3 are based on exact span-and-label agreement only. These scores serve as the main NER evaluation setting in this study.

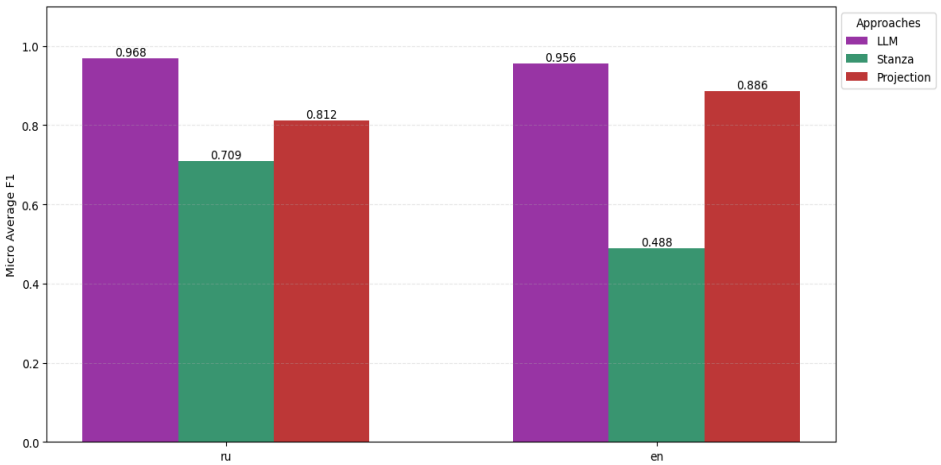

Figure 5. Strict span-level (entity-level) evaluation: F1-score comparison across approaches (Russian and English).

Table 3. Span-level NER performance (strict match): precision, recall, F1, and TP/FP/FN counts.

Metric	LLM (RU)	LLM (EN)	Stanza (RU)	Stanza (EN)	Projection (RU)	Projection (EN)
F1	0.9682	0.956	0.7087	0.4885	0.812	0.8857
Precision	0.9744	0.9383	0.9375	0.6038	1.0	1.0
Recall	0.962	0.9744	0.5696	0.4103	0.6835	0.7949
TP	76	76	45	32	54	62
FP	2	5	3	21	0	0
FN	3	2	34	46	25	16

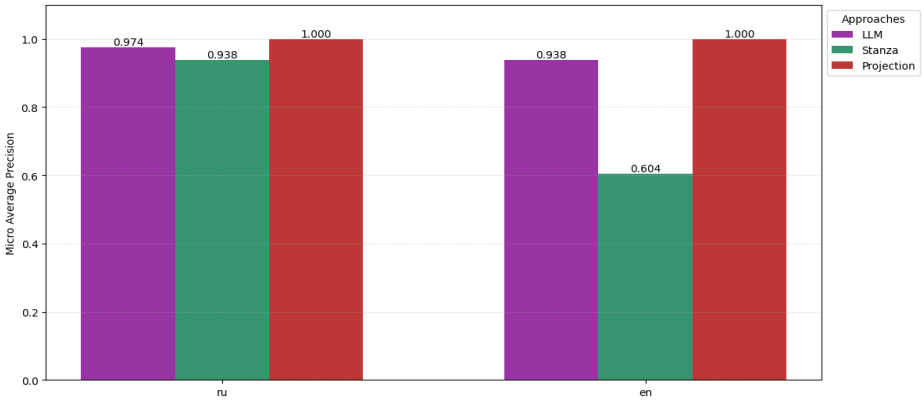

Figure 6. Strict span-level (entity-level) evaluation: precision comparison across approaches (Russian and English).

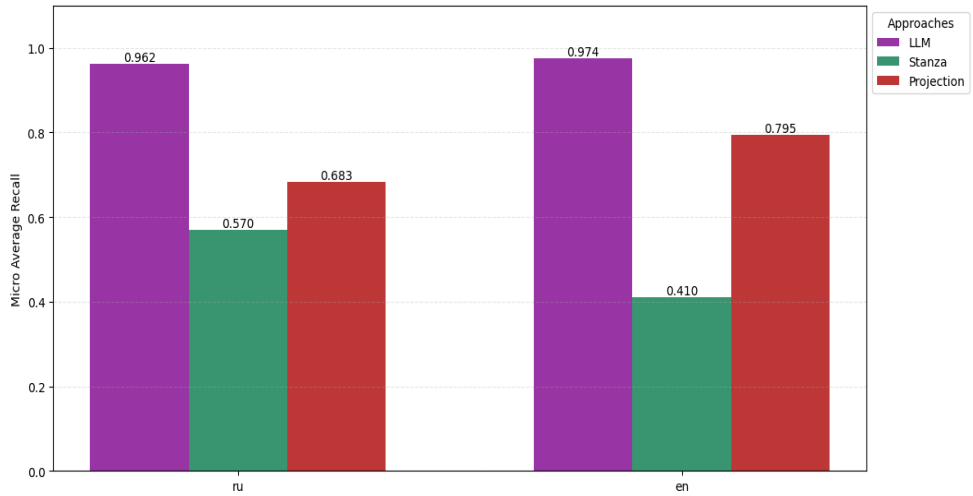

Figure 7. Strict span-level (entity-level) evaluation: recall comparison across approaches (Russian and English).

Several consistent patterns can be observed:

1. LLM-based approach. The LLM pipeline achieves the most balanced performance. It maintains high Precision while also achieving comparatively higher Recall, indicating better preservation of both entity boundaries and entity types.
2. Annotation Projection. Projection typically yields high Precision, as it directly transfers entities from the original markup and is less likely to introduce spurious entities. However, Recall is often lower. This reduction is mainly due to translation-induced alignment issues, including word-order changes, span fragmentation, or morphological variation.

3. Stanza Baseline. The Stanza baseline generally produces the lowest F1-score. Since it performs independent entity extraction on the translated text, it may generate additional entities not present in the reference (increasing FP) and fail to recover some original entities (increasing FN), particularly for multi-word entities and complex syntactic structures.

These strict results provide a rigorous evaluation of each method's ability to transfer annotations accurately rather than merely identifying semantically related entities.

4.3. Entity Quality Considering Semantic Equivalence

In addition to the strict span-level evaluation reported in Table 3, we conduct a separate alternative evaluation that allows semantic equivalence between entities. This setting is not intended to replace the strict evaluation; rather, it is used to examine how much of the error is caused by translation-induced surface-form variation. Therefore, the scores in Table 4 should be interpreted as a complementary analysis under a different matching criterion, not as a direct replacement for the strict results.

Table 4. Entity-level performance with semantic equivalence: precision, recall, F1, and TP/FP/FN counts.

	LLM Approach		**Stanza Approach**		**Annotation Projection Approach**	
	Russian	**English**	**Russian**	**English**	**Russian**	**English**
F1	0.9554	0.8679	0.6772	0.458	0.7819	0.8143
Precision	0.9615	0.8519	0.8958	0.566	0.963	0.9194
Recall	0.9494	0.8846	0.5443	0.3846	0.6582	0.7308
TP	75	69	43	30	52	57
FP	3	12	5	23	2	5
FN	4	9	36	48	27	21

In this evaluation, an entity is counted as a True Positive if one of the following conditions is met:

1. Exact match: the predicted entity span and entity type match the reference exactly;
2. Semantic match: the predicted entity is matched to a reference entity of the same entity type within the same sentence, and the cosine similarity between their embeddings is $\geq$T_ent = 0.85.

As it was mentioned earlier, entity embeddings are computed using the multilingual Sentence-Transformer model sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2. To avoid multiple predictions matching the same reference entity, we enforce one-to-one matching by greedily assigning the highest-scoring pairs.

Note: This semantic evaluation complements strict span-level scoring by reducing penalties for surface-form variation (e.g., abbreviations vs. full forms), while still requiring consistent entity typing.

The corresponding results are presented in Table 4 and Figures 8–10.

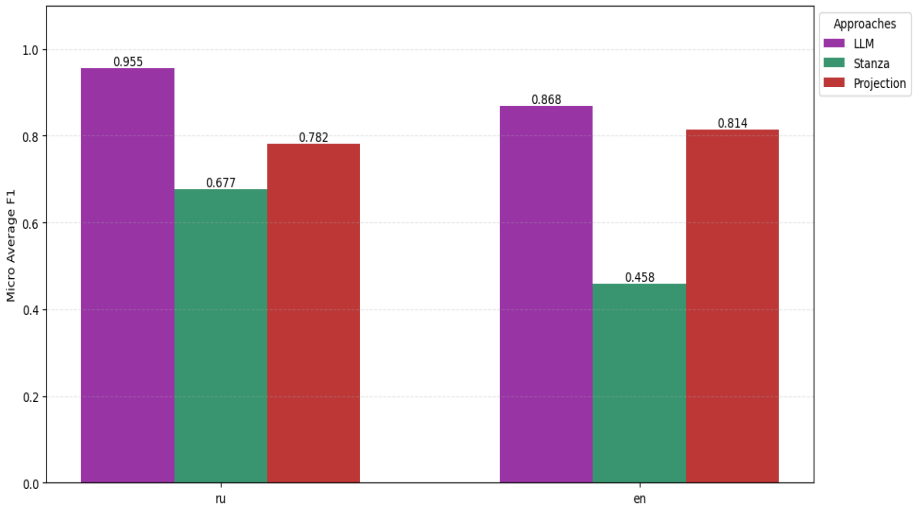

Figure 8. Semantic-equivalence evaluation: F1-score across approaches (Russian and English).

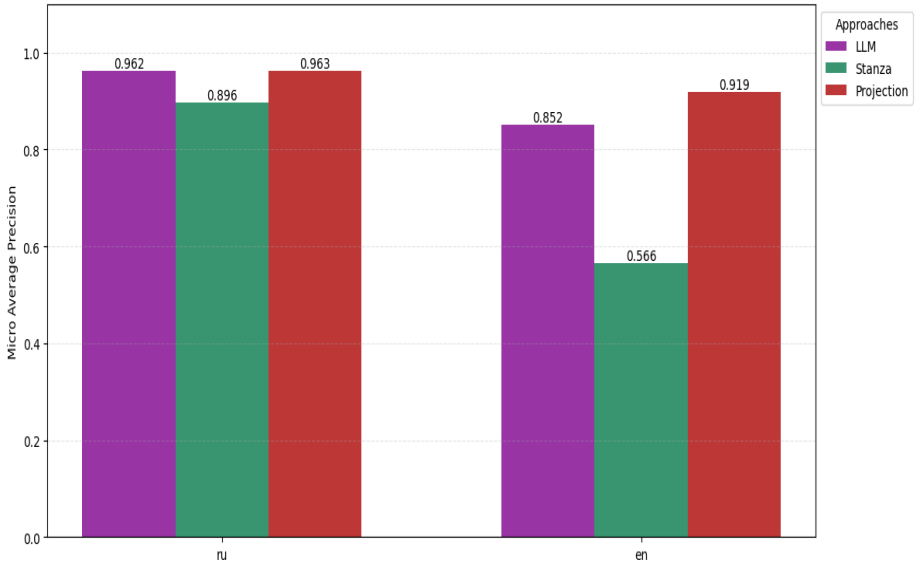

Figure 9. Semantic-equivalence evaluation: Precision across approaches (Russian and English).

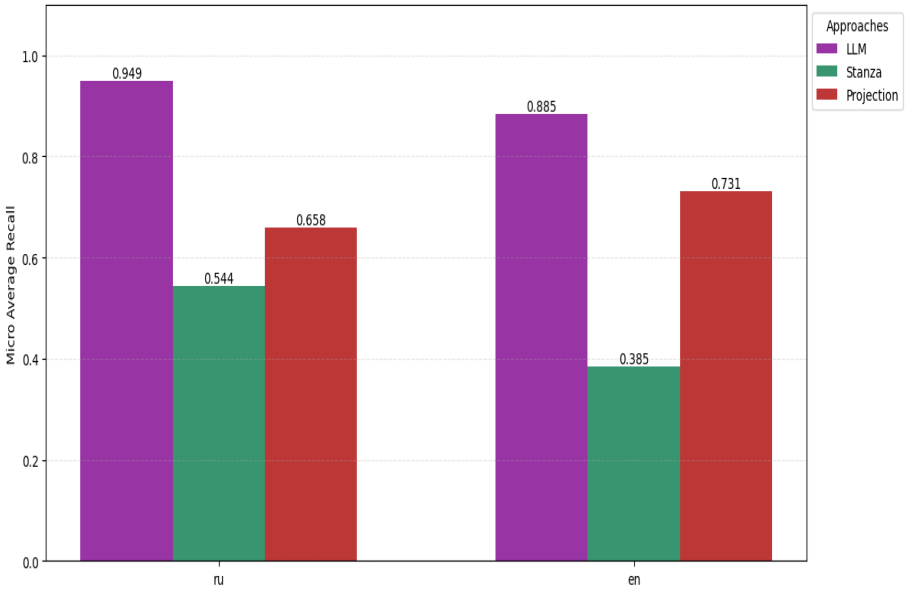

Figure 10. Semantic-equivalence evaluation: Recall across approaches (Russian and English).

The following trends are observed:

1. LLM-based approach. The LLM pipeline benefits most from semantic evaluation. In many cases, it preserves the correct entity in meaning even when the wording differs from the reference annotation.
2. Annotation Projection. The projection shows moderate improvement compared to the strict evaluation. Some string mismatches become correct when semantic equivalence is considered, particularly in cases of alternative translations.
3. Stanza Baseline. Although semantic evaluation slightly improves Stanza's scores, its overall performance remains lower. This is because its primary weakness lies not only in surface variation but also in independent entity extraction, which introduces additional false positives and misses reference entities.

Overall, the semantic evaluation confirms that a portion of strict errors is caused by translation-induced surface variation rather than incorrect entity recognition. However, the relative ranking of methods remains consistent: the LLM-based pipeline performs best across both strict and semantic assessments. For this reason, improvements from Table 3 to Table 4 should be interpreted as the effect of relaxed semantic matching rather than as gains under the original strict protocol.

5. Discussion

This section highlights the main takeaways from the results. Overall, the LLM-based pipeline provides the most stable performance because it better preserves entity boundaries and entity types under cross-lingual transfer. In contrast, annotation projection often loses recall due to translation-induced alignment and boundary shifts, especially for multi-word entities. The Stanza baseline produces more false positives and false negatives because it re-detects entities independently of the source annotation, which can diverge from the transfer-oriented reference standard.

5.1. Key Observations

Overall, the results show that the three approaches yield significantly different results. The LLM-based approach, on average, performs most consistently across both target languages. It more often correctly identifies entity types and better preserves entity boundaries, resulting in more balanced precision and recall values.

Back-translation-based verification helps identify cases where translation changes the intended meaning, which can directly lead to boundary and label errors in NER. Therefore, this step improves the robustness of the LLM-based pipeline.

5.2. Why Methods Produce Different Results

The main advantage of the LLM approach is that the model can take context into account. It does not simply search for word matches but rather attempts to understand the meaning of the sentence. Because of this, LLM is often better at determining whether a word is an entity, its type, and where it begins and ends.

The annotation projection method typically appears more "controlled" because it attempts to transfer only those entities already annotated in the source Uzbek text. This often improves precision, as the method is less likely to add "unnecessary" entities. However, this approach can lose recall if the translation changes word order, if a phrase is shortened or expanded, or if the translation is not literal. In such cases, it is difficult to accurately locate the desired fragment in the translation, and some entities are simply not transferred.

The Stanza baseline is different in nature: it does not transfer the original entities but re-detects entities in the translated text. Therefore, its output can diverge from the reference

annotation used for evaluation. In our setting, this typically increases both false positives (extra entities) and false negatives (missed reference entities).

5.3. Common Errors

When manually reviewing examples, several recurring error types can be identified:

1. Multi-word entities (especially organizations).

Boundary errors often occur here: the method only captures part of the name or captures extra words nearby. This problem occurs with all approaches but is particularly noticeable with projection and Stanza.

2. Translation is not literal but semantic.

Sometimes an entity appears differently in translation: for example, a short name is replaced with an official one or vice versa. In a strict evaluation, this can be considered an error, although the essence is the same. Therefore, an evaluation that takes semantics into account usually shows a more "softer" and more realistic picture of the translated data.

3. The difference between transfer and "search from scratch."

Importantly, Stanza solves a different problem: it does not transfer the source markup but creates new markup in the target language. Therefore, even entities that are correct from a Russian/English perspective may not match the transfer standard. In our setting, this leads to deterioration in the metrics.

The comparison between the strict results (Table 3) and the semantic-equivalence results (Table 4) shows that part of the observed errors are caused by translation-related surface variation rather than complete entity recognition failure. At the same time, Table 4 does not replace the strict evaluation, because it uses a different matching criterion and therefore answers a different analytical question.

5.4. Differences Between Russian and English

The results also show that methods may perform differently for different target languages. For example, in some cases, translation into Russian appears slightly easier than into English. One possible explanation is that translating Uzbek into Russian is often more straightforward, while translating into English may change sentence structure more significantly. When the structure changes more significantly, it becomes more difficult to transfer entity boundaries and match a strict standard.

This observation suggests that structural differences between Uzbek and the target language may affect boundary preservation; however, additional controlled experiments are needed to confirm this hypothesis.

5.5. Limitations and Practical Conclusions

This work has several limitations:

1. All approaches rely on the quality of machine translation. If the translation significantly changes the meaning or structure, the transfer of annotations becomes less accurate.
2. The LLM approach relies on the stability of the response. Sometimes the model may violate the annotation format, so checks and corrections are needed.
3. The semantic similarity analysis depends on the embedding model and the selected thresholds (e.g., T_ent), and results may vary under different settings.
4. The reference set contains 300 sentences for each target language, which already provides a more reliable score than a very small sample, but it still may not cover all text genres and all complex cases.

From a practical standpoint, the LLM approach can be used as a quick way to obtain preliminary annotation for other languages. However, to create a high-quality corpus, it is still useful to apply additional checks and, where possible, partial manual correction, especially for long organization names and ambiguous cases.

6. Conclusions

This paper examined the problem of transferring BIOES annotation for NER from Uzbek to Russian and English. The main challenge to be addressed was the difficulty of quickly obtaining high-quality annotated data for many languages, while manual annotation requires time and expertise.

Three approaches were compared in the study. The first approach is based on LLM, which, according to instructions, simultaneously translates a sentence and outputs BIOES tags. The second approach is annotation transfer via translation (annotation projection), where entities from the source Uzbek text are transferred to the translation. The third approach is a basic version with a ready-made NER model (Stanza), where entities in the translation are newly discovered without using the original annotation.

The experimental results show that the LLM approach generally demonstrates the most consistent performance in both target languages. It better preserves entity boundaries and their types and more often produces correct results even when the translation is not completely literal. At the same time, transferring markup through translation can be useful as a more "controlled" method, but it is sensitive to translation quality and changes in sentence structure. The Stanza approach proved to be the simplest to implement. Still, as a baseline, it demonstrated that ready-made models are not always suitable for markup transfer because they find entities using their own rules and may diverge from the standard.

On the 300-sentence reference sets for each target language, the LLM-based pipeline achieved the highest strict span-level performance (see Table 3). The projection method showed perfect precision but lower recall due to translation-induced boundary mismatches, while the Stanza baseline produced more false positives and false negatives because it re-detects entities independently of the source annotation.

It is also important that the LLM approach's additional verification through back-translation (used as a semantic consistency check) helps identify cases where meaning is significantly distorted, making the pipeline more reliable in practice.

The results also confirm the main contributions stated at the beginning of this paper. First, the proposed LLM-based pipeline provides a practical way to transfer BIOES annotations from Uzbek to Russian and English while preserving entity boundaries and types more reliably than the baseline approaches. Second, the embedding-based semantic analysis helps distinguish strict boundary errors from translation-induced surface-form variation. Third, the comparative evaluation of expert-labeled reference sets shows that the proposed approach is not only effective for the languages studied here but may also be useful as a general strategy for other low-resource languages where manually annotated corpora are limited.

In this sense, the proposed workflow can be viewed as a practical resource-creation strategy for multilingual NER in low-resource settings beyond the Uzbek–Russian–English case. Future work will focus on three directions. First, to increase the volume and diversity of test data across text genres to ensure more robust evaluation. Second, to study entity type errors in more detail and add stricter rules for complex cases (e.g., multi-word entities). Third, to try to combine the best aspects of these methods: using LLM for initial markup and then applying more formal verification and correction steps.

Author Contributions: Conceptualization, V.B. and D.M.; methodology, D.M. and R.M.; software, D.L. and A.K.; validation, D.M., R.M. and A.T.; formal analysis, D.M. and I.A.; investigation, D.M., R.M. and B.I.; resources, B.I. and I.A.; data curation, R.M. and B.I.; writing—original draft preparation, D.M. and R.M.; writing—review and editing, V.B., A.K. and A.T.; visualization, D.L. and A.K.; supervision, V.B. and A.K.; project administration, V.B. and D.M.; funding acquisition, A.K. and R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant No. AP23486904). The research was also conducted within 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.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: We used ChatGPT (OpenAI, GPT-5.2) in two ways: (1) as an LLM in the experimental pipeline to translate sentences and generate candidate cross-lingual BIOES annotations under a fixed prompt template; (2) during manuscript preparation for English language improvement (grammar and spelling). All generated outputs were checked by the authors. All scientific content, analyses, and conclusions were produced and verified by the authors, who take full responsibility for the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. O'Shaughnessy, D. An Overview of Recent Advances in Natural Language Processing and Speech Recognition. *Appl. Sci.* **2026**, *16*, 1122. [CrossRef]
2. Jiang, P.; Cai, X. A Survey of Text-Matching Techniques. *Information* **2024**, *15*, 332. [CrossRef]
3. Seow, W.; Chaturvedi, I.; Hogarth, A.; Mao, R.; Cambria, E. A review of named entity recognition: From learning methods to evaluation metrics. *Artif. Intell. Rev.* **2025**, *58*, 129. [CrossRef]
4. Kühnel, L.; Fluck, J. We are not ready yet: Limitations of state-of-the-art disease named entity recognizers. *J. Biomed. Semant.* **2022**, *13*, 26. [CrossRef] [PubMed]
5. Mengliev, D.; Barakhnin, V.; Abdurakhmonova, N.; Eshkulov, M. Developing named entity recognition algorithms for Uzbek: Dataset insights and implementation. *Data Brief* **2024**, *54*, 110413. [CrossRef] [PubMed]
6. Yang, S.; He, K.; Li, W.; He, Y. CLFF-NER: A Cross-Lingual Feature Fusion Model for Named Entity Recognition in the Traditional Chinese Festival Culture Domain. *Informatics* **2025**, *12*, 136. [CrossRef]
7. Li, J.; Sun, A.; Han, J.; Li, C. A Survey on Deep Learning for Named Entity Recognition. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 50–62. [CrossRef]
8. Komariah, K.S.; Purnomo, A.T.; Satriawan, A.; Hasanuddin, M.O.; Setianingsih, C.; Sin, B.-K. SMPT: A Semi-Supervised Multi-Model Prediction Technique for Food Ingredient Named Entity Recognition (FINER) Dataset Construction. *Informatics* **2023**, *10*, 10. [CrossRef]
9. Mengliev, D.; Barakhnin, V.; Eshkulov, M.; Ibragimoiv, B.; Madirimov, S. A comprehensive dataset and neural network approach for named entity recognition in the uzbek language. *Data Brief* **2025**, *58*, 111249. [CrossRef] [PubMed]
10. Kim, J.; Ko, Y.; Seo, J. Construction of machine-labeled data for improving named entity recognition by transfer learning. *IEEE Access* **2020**, *8*, 59684–59693. [CrossRef]
11. Chiu, J.; Nichols, E. Named Entity Recognition with Bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 357–370. [CrossRef]
12. Elov, B.; Samatboyeva, M. Identifying ner (named entity recognition) objects in Uzbek language texts. *Sci. Innov. Int. Sci. J.* **2023**, *2*, 44–57. [CrossRef]
13. Ji, B.; Liu, R.; Li, S.; Yu, J.; Wu, Q.; Tan, Y.; Wu, J. A hybrid approach for named entity recognition in Chinese electronic medical record. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 64. [CrossRef] [PubMed]
14. Ehrmann, M.; Turchi, M.; Steinberger, R. Building a Multilingual Named Entity-Annotated Corpus Using Annotation Projection. In *Proceedings of RANLP 2011*; Association for Computational Linguistics: Hissar, Bulgaria, 2011; pp. 118–124.

15. MacLean, C.; Cavallucci, D. Assessing fine-tuned NER models with limited data in French: Automating detection of new technologies, technological domains, and startup names in renewable energy. *Mach. Learn. Knowl. Extr.* **2024**, *2024*, 1953–1968. [CrossRef]
16. Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; McClosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of ACL 2014: System Demonstrations*; Association for Computational Linguistics: Baltimore, MD, USA, 2014; pp. 55–60.
17. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186. [CrossRef]
18. Mengliev, D.; Barakhnin, V.; Ibragimov, B.; Eshkulov, M.; Allamov, O.; Shohrux, M.; Khujaev, O.; Rakhimov, B. Development of hybrid approach for named entity recognition in Uzbek language text. *PeerJ Comput. Sci.* **2026**, *12*, e3489. [CrossRef]
19. Mengliev, D.; Barakhnin, V.; Atakhanov, M.; Ibragimov, B.; Eshkulov, M.; Saidov, B. Developing rule-based and gazetteer lists for named entity recognition in Uzbek language: Geographical names. In *Proceedings of the 2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE)*; Institute of Electrical and Electronics Engineers Inc.: Novosibirsk, Russia, 2023; pp. 1500–1504. [CrossRef]
20. Hu, Z.; Li, W.; Yang, H. Named Entity Recognition in Online Medical Consultation Using Deep Learning (BIOES-Y tagging description). *Appl. Sci.* **2025**, *15*, 3033. [CrossRef]
21. Wang, S.; Sun, X.; Li, X.; Ouyang, R.; Wu, F.; Zhang, T.; Li, J.; Wang, G.; Guo, C. Named Entity Recognition via Large Language Models (GPT-NER). In *Findings of NAACL 2025*; Association for Computational Linguistics: Albuquerque, NM, USA, 2025; pp. 4257–4275. [CrossRef]