applied sciences

MDPI

Article

LLM-Assisted Weak Supervision for Low-Resource Kazakh Sequence Labeling: Synthetic Annotation and CRF-Refined NER/POS Models

Aigerim Aitim

Department of Information Systems, International Information Technology University, Manas str. 34/1, Almaty 050000, Kazakhstan; a.aitim@iitu.edu.kz

Abstract

Kazakh sequence labeling is constrained by limited annotated resources, while its agglutinative morphology and productive suffixation increase data sparsity and exacerbate label inconsistency in part-of-speech (POS) tagging and named entity recognition (NER). This paper proposes an LLM-assisted weak supervision framework in which a large language model generates synthetic token-level annotations that are subsequently filtered using confidence-based criteria and combined with a smaller manually verified subset to train Transformer-based sequence taggers with Conditional Random Field (CRF) decoding. The pipeline unifies corpus construction, weak-label generation, quality filtering, word-to-subword alignment, and CRF-refined structured prediction into a reproducible workflow. Experimental results show that contextual encoders and structured decoding provide strong performance for Kazakh POS and NER, while the proposed training design enables efficient convergence with diminishing returns beyond moderate epoch budgets. Error-slice analysis indicates that residual errors are concentrated in rare tokens, morphologically complex long words, longer sentences, and the ORG entity class. Overall, the findings support the use of LLM-assisted weak supervision as a scalable strategy for low-resource Kazakh sequence labeling when synthetic labels are controlled through filtering and refined by structured decoding.

Keywords: Kazakh NLP; low-resource languages; large language models; synthetic annotation; sequence labeling; part-of-speech tagging; named entity recognition; Conditional Random Fields; transformer models

Check for updates

Academic Editors: Francisco De Arriba-Pérez and Silvia García-Méndez

Received: 11 March 2026
Revised: 31 March 2026
Accepted: 6 April 2026
Published: 8 April 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

In low-resource natural language processing scenarios, adaptive and data-efficient techniques are gaining importance to reduce reliance on large manually annotated datasets. Pretrained Transformer encoders (e.g., BERT) and multilingual transfer models (e.g., XLM-R) have improved token-level prediction and cross-lingual generalization in low-resource settings [1,2]. Transfer-based supervision, such as multilingual label propagation, can further alleviate data scarcity for sequence labeling. Weak supervision has emerged as an effective way to augment training data when expert labeling is costly or inaccessible. In sequence labeling tasks like part-of-speech tagging and named entity recognition, this issue is particularly significant for morphologically rich languages, where annotation necessitates both linguistic competence and consistency across intricate token boundaries and inflectional patterns [3]. Notwithstanding the swift advancement of large language

models and pretrained encoders, numerous low-resource language pipelines continue to depend on restricted manually labeled datasets and limited annotation pipelines. In many practical contexts, the same limited labeled samples are routinely utilized for model training, while the extensive corpus of unlabeled text remains mostly underutilized. This situation limits scalability and frequently results in unstable generalization, especially for infrequent, ambiguous, or morphologically intricate forms.

This paper proposes an LLM-assisted weak supervision framework for low-resource Kazakh sequence labeling to overcome this issue. The framework generates weak token-level labels for POS tagging and NER, then refines them through confidence filtering, subword alignment, and CRF-based decoding. The objective of the proposed strategy is to mitigate annotation bottlenecks while maintaining sequence-level consistency and practical model robustness. To address the limitations of purely manual annotation and unstable automatic labeling, the approach integrates weak label generation using large language models with a distinctly interpretable training framework that includes confidence-based filtering, subword-aware alignment, and Conditional Random Field decoding for output refinement.

The framework is evaluated on a proprietary Kazakh news corpus through controlled experiments in POS tagging and named entity recognition. The comparison includes baseline settings with weaker supervision and less structured decoding.

Objectives and Contributions

This paper proposes the creation of a reproducible and annotation-efficient weak supervision system that dynamically enhances usable training material for Kazakh sequence labeling while maintaining the quality of structured predictions. The objectives of this study are as follows:

(1) Design an LLM-assisted pipeline to generate weak token-level labels for Kazakh POS tagging and NER;
(2) Integrate confidence-based filtering, word–subword alignment, and CRF decoding into a unified training workflow;
(3) Quantify the impact of weak supervision via controlled comparisons against gold-only and non-structured baselines;
(4) Analyze convergence behavior and error patterns (rare/frequent tokens, word length, sentence length, and entity classes) to identify remaining failure modes.

The main contribution of this study is the design of a controlled and reproducible weak-supervision pipeline for Kazakh sequence labeling, in which LLM-generated labels are treated as an intermediate supervision source rather than as direct replacements for gold annotation.

The research employs a proprietary Kazakh news corpus and assesses the proposed framework via experiments in POS tagging and named entity recognition, focusing on comparative performance, error characteristics, and sequence-level consistency under limited supervision.

The remainder of this paper is organized as follows. Section 2 reviews related work on low-resource sequence labeling, weak supervision, and structured decoding. Section 3 presents the proposed framework, including corpus construction, weak-label generation and filtering, subword-aware alignment, and the Transformer + CRF model design. Section 4 reports experimental results, including comparative evaluation, convergence analysis, and slice-based error analysis. Section 5 discusses implications, robustness, and practical considerations. Section 6 concludes the paper and outlines future directions.

https://doi.org/10.3390/app16083632

2. Literature Review

Recent advances in pretrained language models and data-centric natural language processing have significantly improved sequence labeling, particularly for tasks such as part-of-speech tagging and named entity recognition. Devlin et al. introduced BERT as a strong bidirectional contextual encoder, demonstrating that large-scale pretraining substantially improves token-level prediction across downstream tasks [1]. Building on this paradigm, Conneau et al. proposed XLM-R, showing that multilingual pretraining makes contextual modeling more accessible to low-resource languages by enabling cross-lingual transfer and stronger performance under limited annotation [2]. As a result, low-resource NLP has increasingly shifted from fully manual annotation toward hybrid strategies that combine pretrained models, weak supervision, and selective human correction.

Research on annotation-efficient sequence labeling can be broadly divided into two major directions: transfer-based supervision and weak supervision. In transfer-based methods, multilingual encoders or projected labels from high-resource languages are used to provide supervision in low-resource settings. For example, Imani, Severini, Jalili Sabet, Yvon, and Schütze showed that graph-based multilingual label propagation can substantially improve low-resource POS tagging by refining projected labels from multiple source languages [3]. In parallel, weak supervision aims to reduce dependence on manually labeled corpora by generating noisy labels from inexpensive supervision sources and then aggregating or denoising them. Ratner, Bach, and colleagues formalized this paradigm in Snorkel, demonstrating that multiple weak labeling functions can be combined into competitive training signals without exhaustive manual annotation [4]. These two strands—cross-lingual transfer and weak supervision—continue to define much of the methodological landscape for low-resource sequence labeling.

At the modeling level, most practical sequence-labeling pipelines still rely on subword tokenization and structured decoding. Kudo and Richardson's SentencePiece remains one of the most widely used approaches for subword segmentation because it reduces out-of-vocabulary errors and allows Transformer models to generalize better to rare lexical forms [5]. However, subword tokenization introduces a persistent difficulty for sequence labeling: labels are typically defined at the word level, whereas the model consumes subword units. This creates non-trivial alignment challenges, especially in morphologically rich and agglutinative languages, where grammatically meaningful suffix chains may be split across several pieces. To preserve sequence-level consistency under such conditions, Conditional Random Fields (CRFs), introduced by Lafferty, McCallum, and Pereira, remain widely used because they explicitly model label transitions and help enforce coherent BIO sequences in NER [6].

More recently, large language models have begun to function not only as end-task predictors, but also as supervision sources. Smith, Fries, Hancock, and Bach showed in Language Models in the Loop that prompted language models can be embedded into a weak-supervision framework as labeling functions, yielding substantial gains over zero-shot baselines while remaining competitive with systems based on hand-engineered weak rules [7]. Extending this idea, Su, Yu, Zhang, and Bach proposed using LLMs not only to generate prompted labels, but also to estimate dependency structure among prompted labeling functions, thereby improving the denoising stage of weak supervision and strengthening the final synthetic labels [8]. These studies are especially relevant because they suggest that LLMs are most effective in low-resource sequence labeling when they are used to generate and structure supervision, rather than simply to produce direct predictions. More broadly, recent evaluations of LLM-based agents suggest that strong general capabilities do not guarantee uniformly reliable behavior across constrained task settings, which further motivates using LLM outputs as controlled supervision signals rather than as fully trusted

final predictions [9]. Recent LLM-based NER studies also suggest that raw prompting can be effective, but often remains sensitive to output formatting, label consistency, and task-specific control, especially when prompts are used directly as pseudo-label generators rather than as components of a constrained annotation pipeline.

At the same time, the literature reveals several unresolved methodological debates. One debate concerns the extent to which strong pretrained encoders can replace explicit linguistic modeling. One position argues that sufficiently strong contextual encoders can absorb much of the burden of morphology and sequence modeling, reducing the need for explicit structure-aware preprocessing. A competing view maintains that low-resource settings still benefit from explicit structural constraints, particularly when supervision is noisy. A related debate concerns tokenization itself. Clark, Garrette, Turc, and Wieting showed with CANINE that tokenization-free encoders can avoid the brittleness of subword vocabularies by operating directly on character sequences [10]. Similarly, Xue et al. proposed ByT5, showing that byte-level pretraining can remove tokenization altogether while remaining competitive across multilingual tasks [11]. These works support the view that tokenization itself may be a source of unnecessary complexity for morphologically rich languages.

Nonetheless, tokenization-free designs involve trade-offs regarding sequence length, computational expense, and optimization efficacy, and they have not supplanted subword-based frameworks in most practical low-resource applications. As a result, there is still no consensus on whether low-resource sequence labeling should emphasize more robust subword encoders, tokenization-free designs, or subword pipelines enhanced with structured decoding.

In Kazakhstan, the resource landscape has improved; nonetheless, significant limitations persist. Yeshpanov, Khassanov, and Varol established KazNERD as a specialized dataset for Kazakh named entity recognition, including an extensive benchmark with annotation protocols to enhance repeatability and comparability among systems. This was a major advancement for Kazakh NER; nonetheless, it does not resolve the overarching issue because extensive, varied, token-level corpora for numerous sequence-labeling tasks are still lacking [12]. Practically, a substantial gap persists between the volume of accessible Kazakh material and the quantity of text that is adequately annotated for effective supervised learning across several domains [13].

Positioning vs. prior weak-supervision and LLM-annotation pipelines. Classical weak supervision frameworks (e.g., Snorkel-style pipelines) treat supervision as combinations of labeling functions and focus on denoising/aggregation at the label-model stage [14]. In contrast, recent LLM-based annotation pipelines often use prompted outputs as direct pseudo-labels and rely on minimal post-processing. The proposed framework sits between these extremes: it uses an LLM only as a candidate label generator, but explicitly controls label quality through (i) schema-constrained prompting and validator-based rejection, (ii) self-consistency confidence from repeated generations, (iii) threshold-based filtering to trade coverage for noise, and (iv) CRF decoding to enforce sequence-level constraints under BIO tagging. This design aims to attribute gains to pipeline controls rather than to raw LLM strength.

Relative to Snorkel-like pipelines, the proposed framework does not rely on manually designed labeling functions or a separate label model for denoising. Relative to direct LLM pseudo-labeling approaches, it does not treat prompted outputs as training-ready annotations. Instead, it introduces a constrained intermediate supervision layer in which candidate labels are validated, aggregated across repeated generations, filtered by confidence, aligned to subword tokenization, and only then passed to a structured downstream model with CRF decoding. The novelty of the present work lies in this controlled integra-

tion of LLM-based weak labeling, explicit quality control, and sequence-level refinement for low-resource Kazakh sequence labeling.

Taken together, the literature suggests a methodological trade-off. Snorkel-like systems offer explicit control and denoising structure, but typically depend on manually designed supervision sources. Direct LLM pseudo-labeling offers broader coverage and lower rule-engineering cost, but often provides weaker guarantees of output consistency and annotation reliability. The present study is positioned as a middle-ground design: it uses an LLM to scale supervision, but introduces explicit control mechanisms to recover some of the reliability typically associated with more structured weak-supervision pipelines.

This trade-off is central to the present study: rather than maximizing coverage alone or relying on manual weak-rule engineering, the proposed framework prioritizes controllable supervision quality under low-resource conditions.

Research Gap. Taken together, prior work suggests a shift from fully manual labeling toward annotation-efficient pipelines that combine multilingual contextual encoders, weak supervision, and structured decoding [15–18]. However, existing studies typically emphasize only one component at a time (transfer-based supervision, weak-label aggregation, or tokenization-free modeling) rather than integrating LLM-generated weak supervision, explicit quality filtering, subword-aware alignment, and CRF-based refinement within one reproducible sequence-labeling workflow, particularly for morphologically rich, low-resource languages such as Kazakh.

3. Materials and Methods

The proposed method provides a systematic framework for weak supervision in low-resource Kazakh sequence labeling, emphasizing part-of-speech (POS) tagging and named entity recognition (NER). The proposed method uses large language models as synthetic label generators within a controlled training workflow, in contrast to traditional pipelines that rely predominantly on manual annotation or direct zero-shot prompting. The approach integrates corpus construction, weak-label generation, confidence-based filtering, selective expert correction, subword-aware alignment, Transformer-based contextual encoding, and Conditional Random Field (CRF) decoding into a singular reproducible workflow. The method seeks to minimize annotation cost while maintaining sequence-level consistency and practical robustness in low-resource environments.

3.1. Conceptual Framework

The primary concept of the proposed approach is to convert unlabeled Kazakh text into functional training data via a phased weak-supervision methodology. Instead of relying on an LLM as the final prediction model, the system uses the LLM to produce candidate token-level labels that are subsequently filtered and refined before training a dedicated sequence-labeling model. This design distinguishes label generation from final inference, allowing the final POS/NER model to remain compact, reproducible, and suitable for deployment. The framework is founded on three principles. First, weak supervision must expand training coverage without being accepted uncritically. Second, noisy synthetic labels should be incorporated only with explicit quality control rather than being blended into training without constraint [19,20]. Third, sequence-level consistency must be reinstated during the final decoding phase via CRF-based structured prediction. Together, these principles provide an annotation-efficient yet regulated alternative to entirely manual labeling.

3.2. Corpus Construction and Data Preparation

Experiments were conducted using a proprietary Kazakh news corpus gathered via a web-crawling and extraction pipeline. Following the processes of cleaning, language filtering,

and deduplication, the corpus comprises approximately 3500 to 3800 Kazakh-language articles, amounting to roughly 2.5-million-word tokens and approximately 122,000 sentences. The corpus undergoes normalization via Unicode cleanliness, elimination of boilerplate content, normalizing of whitespace, and standardization of punctuation. Near-duplicate documents are eliminated at the document level to decrease redundancy and mitigate train-test leakage.

A portion of the corpus is manually validated and categorized into training, development, and testing segments at the document level for supervised calibration and assessment. This partitioning technique mitigates leakage from recurring entities or nearly similar news pieces across segments and maintains a realistic evaluation framework for low-resource sequence tagging.

3.3. LLM-Assisted Weak Label Generation

Weak labels were generated using Qwen2.5-72B-Instruct in zero-shot mode with schema-constrained prompting. The same model was used for both POS tagging and NER weak-label generation in order to keep the annotation conditions consistent across tasks. Decoding parameters were fixed for all runs (temperature = 0.2, top_p = 1.0, max_tokens = 256) to reduce stochastic variation and to encourage stable structured outputs. Each sentence was annotated K = 5 times, and the final weak label was obtained by majority aggregation with confidence scoring based on self-consistency (Section 3.3.2). Invalid generations, including token-count mismatches, labels outside the predefined inventory, malformed BIO transitions, and empty outputs, were rejected before confidence filtering. Weak labeling was applied only to the training split; development and test sets were never exposed to the LLM in order to avoid data leakage.

Computational cost and scalability. Weak labeling is the dominant offline cost of the pipeline. In our setting, the training split contains 87,600 sentences and each sentence is labeled K = 5 times, resulting in 438,000 LLM generations. Based on corpus statistics (≈2.5M word tokens over ≈122k sentences), the average sentence length is ≈20.5 word tokens, so the output length is on the same order (one tag per token). Overall complexity scales linearly as O(N·K) with the number of sentences and repetitions. We report the total number of generations in the text, while acceptance rate and retained supervision volume are reflected in the threshold-sensitivity analysis; to support reproducibility, the released scripts include token/count logging and wall-clock timing hooks for any deployment environment.

Qwen2.5-72B-Instruct was selected as the weak-label generator because this study requires stable schema-constrained token-level outputs rather than open-ended generation. In our setting, the model must return exactly one label per whitespace token, obey a fixed label inventory, and avoid malformed BIO transitions. We therefore prioritized three practical criteria: (i) formatting reliability under constrained prompting, (ii) token-level consistency under repeated sampling, and (iii) multilingual instruction-following with English prompts and Kazakh inputs. In preliminary annotation trials on a small development subset, Qwen2.5 produced fewer malformed outputs and more stable token-count-preserving sequences than several alternative instruction-tuned models, which made it a suitable generator for the weak-supervision stage. We did not select the generator on the basis of broad general-purpose benchmark strength alone, but on practical suitability for constrained token-level annotation, where formatting stability was more critical than open-ended generation quality. Importantly, the contribution of this study is not the choice of one particular LLM, but the controlled pipeline in which weak labels are validated, filtered, and refined before downstream training.

This interpretation is also supported by the ablation results, where performance improves progressively from unfiltered weak supervision to filtered weak supervision and

then to CRF-refined decoding, indicating that downstream gains arise from the supervision-control pipeline rather than from raw LLM outputs alone.

All prompts were written in English, while the input sentences remained in Kazakh, because English task instructions produced more stable formatting without altering the lexical content of the target-language input. The annotation pipeline processed 87,600 training sentences. Because the weak-label generation stage was executed across a non-uniform experimental environment, we do not report exact wall-clock throughput or monetary cost as benchmark figures. Instead, we report the number of sentences processed, the number of repeated generations per sentence, and the total number of generations, which provide a hardware-independent view of annotation scale. Future work should include controlled runtime and cost benchmarking across deployment settings.

3.3.1. Prompt Schema and Output Constraints

The LLM was instructed to return token-level labels in a strict machine-readable format to minimize free-form generation. For POS tagging, the model outputs exactly one POS tag per whitespace-separated token using a fixed tag inventory. For NER, the model outputs BIO tags over the entity types PER, LOC, and ORG. Any output that violates the required format, including missing labels, labels outside the target inventory, malformed BIO transitions, or explanatory text, is rejected during quality control [21–24].

Representative prompts were as follows.

POS prompt:

You are a linguistic annotation system. Assign exactly one POS tag to each whitespace-separated token in the Kazakh sentence. Use only the predefined tag inventory: {NOUN, VERB, ADJ, ADV, PRON, NUM, ADP, CCONJ, SCONJ, PART, INTJ, PROPN, AUX, PUNCT, X}. Return only the sequence of tags, separated by single spaces. Do not add explanations.

Sentence: <Kazakh sentence>

NER prompt:

You are a sequence labeling system. Assign BIO tags to each whitespace-separated token in the Kazakh sentence using only the entity types {PER, LOC, ORG}. Use the BIO inventory {B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG, O}. Return only the sequence of tags, separated by single spaces. Do not add explanations.

Sentence: <Kazakh sentence>

A candidate output was accepted only if: (i) the number of predicted labels exactly matched the number of whitespace tokens; (ii) all labels belonged to the permitted inventory; (iii) for NER, BIO transitions satisfied basic sequence constraints; and (iv) the output contained no explanatory text, punctuation artifacts, or merged labels. Sequences failing any validation step were discarded before aggregation.

Prompt and decoding sensitivity controls. To reduce prompt-induced variance, we fix the prompt templates (reported verbatim), use a strict validator to reject malformed outputs, and keep decoding parameters constant (temperature = 0.2, top_p = 1.0, max_tokens = 256). Stability is further improved via K = 5 repeated generations and majority aggregation, which reduces dependence on any single stochastic completion. The released repository includes the exact prompt templates and validation rules to ensure reproducibility. These controls do not eliminate prompt sensitivity entirely, but they reduce its practical impact by constraining the output space, rejecting malformed generations, and replacing single-sample supervision with repeated-sample aggregation. The weak-labeling stage should therefore be interpreted as controlled stochastic annotation rather than as deterministic prompting. In this sense, prompt design is treated as a controlled component of the annotation protocol rather than as an unbounded source of task variation. This makes

control over format validity, label consistency, and downstream integration particularly important when LLMs are used as annotation engines rather than final predictors.

3.3.2. Confidence Estimation

The same sentence $x = (w_1, \ldots, w_T)$ is labeled $K = 5$ times by the LLM. Let $y_t^{(k)}$ denote the label predicted for token k-th generation. The final weak label $\tilde{y}_t$ is obtained by majority vote over the K outputs.

For POS tagging, token-level agreement is defined as the vote proportion of the majority label at position t. The instance-level POS confidence is the mean token agreement across the sentence.

For NER, confidence is computed analogously over the BIO label sequence.

A weakly labeled sentence is retained only if $c(x) \geq \tau$, where τ is the confidence threshold Equations (1)–(5):

$$a_t^{POS} = \frac{1}{K} \max_{l \in \mathcal{L}_{POS}} \sum_{k=1}^{K} 1\left(y_t^{(k)} = l\right) \tag{1}$$

$$c_{POS}(x) = \frac{1}{T} \sum_{t=1}^{T} a_t^{POS} \tag{2}$$

$$a_t^{NER} = \frac{1}{K} \max_{l \in \mathcal{L}_{NER}} \sum_{k=1}^{K} 1\left(y_t^{(k)} = l\right) \tag{3}$$

$$c_{NER}(x) = \frac{1}{T} \sum_{t=1}^{T} a_t^{NER} \tag{4}$$

$$c(x) \geq \tau \tag{5}$$

The same sentence is labeled K = 5 times; the final weak label is obtained by majority vote. Instance-level confidence $c \in [0, 1]$ is defined as the mean vote proportion over token labels for POS and over exact-match entity spans for NER, averaged across the sequence [25–29].

3.4. Weak-Label Filtering and Supervision Formalization

Let D_{gold} denote the manually curated gold-labeled dataset and let D_{weak} denote the filtered weakly labeled dataset generated by the LLM. Each instance in D_{weak} is represented as (x, ỹ, c), where x is the sentence, ỹ is the synthetic label sequence, and c is the confidence score.

The overall training objective combines supervision from manually verified labels and filtered weak supervision Equation (6):

$$\mathcal{L} = \mathcal{L}_{gold} + \lambda_w \mathcal{L}_{weak}, \tag{6}$$

where λ_w controls the contribution of weak supervision.

For the manually curated gold-labeled dataset, the supervised loss is defined as Equation (7):

$$\mathcal{L}_{gold} = \sum_{(x,\, y) \in D_{gold}} -\log P(y|x), \tag{7}$$

where x is the input sequence and y is the corresponding gold label sequence.

For the filtered weakly labeled dataset D_{weak}, each weak instance is represented as $(x, \tilde{y}, c)$, where $\tilde{y}$ the synthetic label sequence generated by the LLM and $c \in [0, 1]$ is its confidence score. The weak-supervision loss is defined as Equation (8):

$$\mathcal{L}_{weak} = \sum_{(x,\tilde{y},c) \in D_{weak}} c \cdot \left(-\log P\left(\tilde{y} \middle| x\right)\right) \tag{8}$$

This formulation weights the contribution of each weak label by its estimated reliability.

The formulation guarantees that synthetic labels contribute in accordance with their assessed reliability. High-confidence weakly labeled instances exert a stronger influence on training, whereas low-confidence labels are either diminished in weight or omitted. The outcome is a transparent supervisory framework wherein weak labels function as weighted supplementary supervision instead of direct substitutes for gold-standard data.

3.5. Subword Tokenization and Word–Subword Alignment

Let an input sentence be represented as a sequence of word tokens Equation (9):

$$W = (w_1, w_2, \ldots, w_n). \tag{9}$$

Each word w_i is segmented into a sequence of subword units Equation (10):

$$w_i \rightarrow (s_{i,1}, s_{i,2}, \ldots, s_{i,k_i}), \tag{10}$$

where k_i is the number of subwords corresponding to w_i.

Under the first-subword supervision strategy, the word-level label y_i is assigned only to the first subword $s_{i,1}$ Equation (11):

$$\hat{y}_{i,1} = y_i. \tag{11}$$

The remaining subwords are excluded from the loss computation Equation (12):

$$\hat{y}_{i,j} = \perp, \qquad j = 2, \ldots, k_i, \tag{12}$$

where $\perp$ denotes an ignore label (e.g., -100 in practical implementations).

The ultimate sequence-labeling model employs a Transformer encoder and consequently functions on subword units. Nonetheless, POS and NER labels are delineated at the lexical level. The framework employs first-subword supervision to harmonize these two levels. The alignment process is particularly crucial for Kazakh because of its agglutinative morphology, which commonly generates lengthy words with extensive suffix chains that are often divided into several subword segments.

An illustrative alignment example is shown in Table 1. In agglutinative Kazakh, a single orthographic word may contain a lexical root followed by several suffixes. Because the sequence-labeling target is defined at the word level, the proposed framework assigns the supervision signal only to the first subword piece, while the remaining subwords are masked out from the loss. This strategy preserves label consistency without forcing duplicate supervision over suffix fragments created by the tokenizer.

Table 1. Illustrative example of word-to-subword alignment for a morphologically complex Kazakh word.

Surface Word	Morphemic Analysis	English Gloss	Illustrative Subword Split	Word-Level Label	Supervised Subword	Remaining Subwords
Университеттерімізден consists of the stem университет (‘university’), the plural suffix -тер, the first-person plural possessive suffix -іміз (‘our’), and the ablative suffix -ден (‘from’), yielding the meaning ‘from our universities’.	университет + -тер + -іміз + -ден = = university + plural + our + from = “from our universities”	university + PL + POSS.1PL + ABL (“from our universities”)	университет/тер/іміз/ден	NOUN	университет → NOUN	тер, іміз, ден → IGNORE

Table 1. *Cont.*

Surface Word	Morphemic Analysis	English Gloss	Illustrative Subword Split	Word-Level Label	Supervised Subword	Remaining Subwords
Қазақстаннан consists of Қазақстан ('Kazakhstan') and the ablative suffix -нан, meaning 'from Kazakhstan'."	Қазақстан + -нан = Kazakhstan + from = "from Kazakhstan"	Kazakhstan + ABL ("from Kazakhstan")	Қазақстан/нан	B-LOC	Қазақстан → B-LOC	нан → IGNORE

The subword split is illustrative and tokenizer-dependent. Under first-subword supervision, the word-level label is assigned only to the first subword piece, while all remaining subwords are masked with IGNORE and do not contribute to the loss.

3.6. Contextual Encoding and CRF-Based Sequence Decoding

For each supervised position t, the Transformer encoder produces a contextual representation Equation (13):

$$h_t \in \mathbb{R}^d. \tag{13}$$

If additional morphology-aware features m_t are used, they are fused with the contextual representation as follows Equation (14):

$$\tilde{h_t} = W_f[h_t; m_t] + b_f, \tag{14}$$

where $[h_t; m_t]$ denotes vector concatenation, W_f and b_f are trainable parameters.

A linear output layer converts the final token representation into emission scores over the label set $\mathcal{Y}$ Equation (15):

$$e_t = W_e\tilde{h_t} + b_e, \tag{15}$$

where $e_t \in \mathbb{R}^{|\mathcal{Y}|}$, and $|\mathcal{Y}|$ is the number of possible labels.

For an input sequence and a label sequence $y = (y_1, \ldots, \mathcal{Y}_T)$, the linear-chain CRF assigns the following sequence score Equation (16):

$$s(x, y) = \sum_{t=1}^{T} \left(e_t[y_t] + A_{y_{t-1}, y_t}\right), \tag{16}$$

where is the trainable transition matrix, and A_{y_{t-1}, y_t} is the transition score from label y_{t-1} to label y_t.

After subword alignment, the sentence is encoded with a pretrained Transformer backbone (e.g., Kazakh-adapted or multilingual). For each supervised position t, the encoder produces a contextual vector h_t. A linear classification layer converts these vectors into emission scores e_t over the target label set.

The CRF defines the conditional probability of a label sequence as Equation (17):

$$P\ (y|x) = \frac{\exp(s(x, y))}{\sum_{y' \in \mathcal{Y}(x)} \exp(s(x, y'))}, \tag{17}$$

where $\mathcal{Y}(x)$ denotes the set of all valid label sequences for the input x.

Training of the sequence labeling model is performed by minimizing the negative log-likelihood Equation (18):

$$\mathcal{L}_{CRF} = -logP(y|x). \tag{18}$$

For weak supervision, the same objective is applied to $\tilde{y}$ scaled by the confidence score c.

During inference, the most probable label sequence is obtained via Viterbi decoding Equation (19):

$$y^* = \underset{y \in \mathcal{Y}(x)}{\text{argmax}} s(x, y). \tag{19}$$

This step ensures globally optimal decoding under the learned transition constraints.

3.7. Implementation Procedure

The proposed weak-supervision framework consists of the following steps:

(1) Corpus initialization: a cleaned and deduplicated Kazakh corpus is prepared from crawled news data.
(2) Weak-label generation: the LLM produces candidate POS and NER labels for unlabeled sentences.
(3) Confidence filtering: weakly labeled sentences are accepted, rejected, or routed for manual review.
(4) Gold subset preparation: a smaller expert-verified subset is prepared for calibration, validation, and final testing.
(5) Tokenization and alignment: sentences are tokenized into subwords, and labels are aligned using first-subword supervision.
(6) Model training: the encoder and CRF are trained using the joint objective over gold and filtered weak supervision.
(7) Evaluation: performance is measured using POS accuracy and entity-level precision, recall, and F1 for NER.

This stepwise procedure ensures that the framework remains transparent and reproducible, while minimizing the need for ad hoc manual adjustments during model development.

3.8. Implementation Reproducibility

All experiments are executed with fixed random seeds and a consistent preprocessing/alignment pipeline across runs. Models are trained using the same optimizer settings and evaluated with the same word-level POS accuracy and entity-level NER micro-F1 (exact match) protocol. The codebase logs configuration files and training metadata to ensure reproducibility. Where redistribution of raw news texts is restricted, we provide derived artifacts (label schemas, preprocessing scripts, and split manifests) to support reconstruction of the dataset.

In addition to single-run point estimates, we assess robustness under sentence-level test-set resampling using paired bootstrap analysis (B = 5000). For each comparison, sentences from the held-out test set are resampled with replacement, and POS accuracy and NER micro-F1 are recomputed to estimate 95% confidence intervals for both absolute scores and paired differences. This analysis is intended to quantify sensitivity to test-set sampling variation. Because the present revision focuses on test-set resampling rather than repeated end-to-end retraining, broader multi-seed variance analysis is left for future work. Accordingly, this analysis should be interpreted as evidence of stability under test-set resampling rather than as a full characterization of end-to-end retraining variance across random seeds.

3.9. Integration of LLM Supervision, Structured Prediction, and Kazakh Morphology

The proposed framework is based on the integration of three complementary elements. The LLM offers scalable weak supervision, hence diminishing the expenses associated with manual annotation. Subword-aware alignment ensures interoperability with Transformer encoders while maintaining word-level supervision. CRF decoding implements structured prediction, which is especially crucial when weak labels exhibit local inconsistencies.

This integration is especially important for Kazakh, where agglutinative morphology increases lexical sparsity and tokenization ambiguity [30]. Weak supervision can enhance coverage; however, in the absence of filtering and structured decoding, it may also exacerbate label noise. The current approach treats LLM results as an intermediate supervisory resource that requires constraints and refinement, rather than as direct predictions for the downstream task.

3.10. Novelty of the Proposed Approach

The main contribution of the proposed method is not simply the use of an LLM for annotation, but the design of a controlled weak-supervision architecture in which LLM-generated labels are treated as provisional supervision and are subsequently validated, confidence-filtered, aligned to subword tokenization, and refined through CRF-based structured decoding. In contrast to prior approaches that emphasize either weak-label aggregation alone or direct LLM pseudo-labeling, the present framework integrates these control mechanisms into a single reproducible workflow for low-resource sequence labeling in Table 2.

Table 2. Positioning vs. prior pipelines.

Pipeline Family	Supervision Source	Quality Control	Structured Decoding	Typical Failure Mode
Snorkel-like	labeling functions	label model	optional	LF mismatch/coverage
LLM pseudo-labeling	prompted labels	minimal	rare	format drift/span noise
Ours	prompted labels + K-sampling	validator + confidence + τ	CRF	residual boundary/type errors

Accordingly, the novelty of the present work lies not in using an LLM alone, but in combining LLM-generated weak labels with validator-based rejection, self-consistency filtering, subword-aware alignment, and CRF-based sequence refinement within one reproducible architecture.

The framework is designed to emphasize interpretability, reproducibility, and practical implementation. Every phase of the pipeline—synthetic label generation, filtering, alignment, training, and structured decoding—is distinctly delineated and may be examined autonomously. This makes the strategy suitable for both empirical assessment and practical application to additional low-resource, morphologically complex languages.

4. Experiments and Results

This section assesses the proposed weak supervision framework from three interrelated perspectives: (i) core task efficacy, (ii) structured error patterns, and (iii) training dynamics across different optimization budgets. The given findings are derived from the in-house corpus Kazakh news corpus outlined in Section 3, consisting of around 3500 to 3800 articles, around 2.5-million-word tokens, and approximately 122,000 sentences following cleaning and deduplication. The investigation is structured around comparative benchmarking, class-level NER performance, slice-based resilience, and convergence efficiency.

4.1. Comparative Performance on POS Tagging and NER

The comparison findings indicate that contextual pretraining and structured decoding significantly enhance performance compared to basic baselines. The most robust non-pretrained baseline for POS tagging, utilizing BiLSTM + CRF, achieves an accuracy of 90.5%, whereas mBERT + CRF enhances this to 96.8%, demonstrating a significant advantage from

contextual multilingual pretraining. The optimal architecture (KazBERT + BiLSTM + CRF) achieves an accuracy of 97.5%, indicating that Kazakh-adapted representations enhance robustness for morphologically intricate and sparse forms.

A similar tendency is evident in named entity recognition. In the comparative analysis, KazBERT + CRF achieves an F1 score of 89.4%, surpassing the mBERT-only baseline (84.6%), topping Flair (87.1%), and significantly outperforming a spaCy multilingual pipeline (74.2%). In the final integrated system evaluation, the integrated NER module achieves a micro-F1 score of 95.0%, validating the robustness of the entire pipeline in end-to-end applications.

The diagram in Figure 1 encapsulates the optimization and performance trends of the proposed model during the training process. The left panel illustrates that the negative log-likelihood of the CRF declines significantly during the initial epochs and thereafter stabilizes, signifying rapid convergence of structured decoding. The central panel indicates that POS accuracy increases rapidly and approaches a near-plateau after roughly five epochs, with only marginal improvements thereafter. The right panel illustrates a comparable pattern for NER micro-F1, which consistently improves and nears saturation by the concluding epochs. The largest optimization changes occur in the early epochs, after which both POS accuracy and NER micro-F1 improve only marginally, indicating that the model reaches a stable regime relatively quickly.

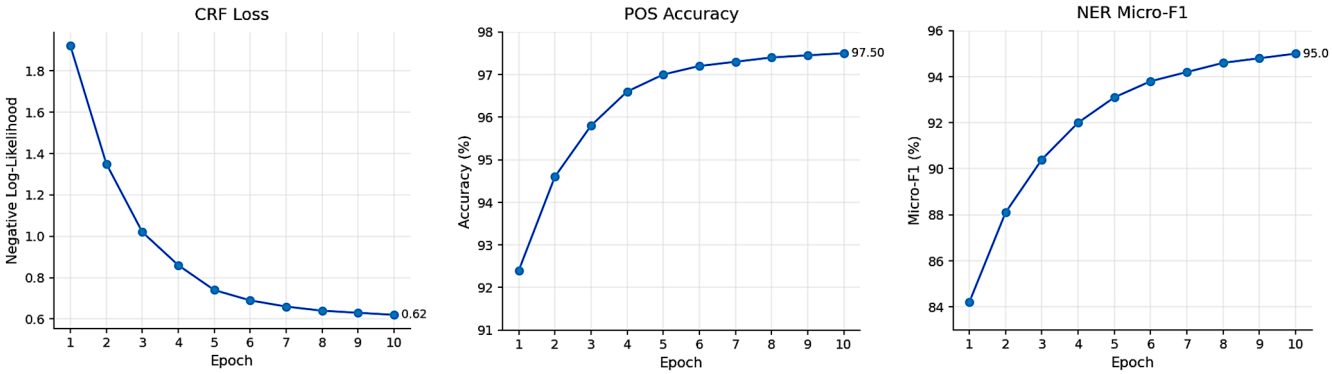


Figure 1. Training dynamics across 10 epochs: (**a**) CRF negative log-likelihood; (**b**) POS accuracy; (**c**) NER micro-F1.

The CRF loss decreases sharply during the first three epochs and then stabilizes, indicating rapid early convergence. POS accuracy improves quickly and approaches a near-plateau by approximately epoch 5, while NER micro-F1 continues to rise more gradually before saturating toward the final epochs [31,32]. Overall, the curves indicate stable optimization and diminishing returns beyond a moderate training budget.

4.2. Effect of LLM-Assisted Weak Supervision

To isolate the contribution of weak supervision, we compare gold-only training against progressively enriched supervision: (i) gold-only, (ii) gold + unfiltered weak supervision, (iii) gold + filtered weak supervision $c \geq \tau$, and (iv) gold + filtered weak supervision with CRF decoding [33]. This controlled comparison quantifies whether LLM-generated labels provide usable signal beyond the manually verified subset and whether structured decoding mitigates weak-label noise [34]. All weak labels are generated on the training split only; development and test sets are excluded to prevent leakage.

Table 3 isolates the contribution of weak supervision. Unfiltered weak supervision provides limited gains due to noise, while confidence filtering yields consistent improvements.

Adding CRF decoding further improves sequence-level coherence under BIO constraints, producing the best overall performance [35].

Table 3. Weak supervision ablation (gold-only vs. weak supervision).

Training Setting	Weak Instances Used	POS Accuracy (%)	ΔPOS vs. Gold	NER Micro-F1 (%)	ΔNER vs. Gold
Gold-only	0	96.2	—	92.0	—
Gold + weak (unfiltered)	87,600	96.4	+0.2	92.6	+0.6
Gold + weak (filtered, $\tau = 0.70$)	52,560	97.1	+0.9	94.6	+2.6
Gold + weak (filtered, $\tau = 0.70$) + CRF	52,560	97.5	+1.3	95.0	+3.0

Bootstrap confidence intervals and significance. Table 4 reports 95% paired-bootstrap confidence intervals and paired-bootstrap significance tests for the main comparisons (C3 vs. C0; C2 vs. C0; C3 vs. C2). This analysis complements single-run point estimates by quantifying the stability of improvements under sentence-level resampling of the test set. B = 5000 sentence-level resamples; Δ is reported in percentage points (pp); p_boot is the fraction of bootstrap samples where $\Delta \leq 0$. These results strengthen the empirical validity of the main comparisons, although they should be interpreted as robustness under resampling rather than as a substitute for full multi-seed retraining analysis.

Table 4. Paired bootstrap robustness on the test set (B = 5000).

(A) Point estimate with 95% CI				
Setting		**POS Acc. (%)**	**NER Micro-F1 (%)**	
C0 Gold-only		96.2 [95.8, 96.6]	92.0 [91.1, 92.9]	
C1 + Weak (unfiltered)		96.4 [96.0, 96.8]	92.6 [91.8, 93.4]	
C2 + Weak (filtered, $\tau = 0.70$)		97.1 [96.8, 97.4]	94.6 [93.9, 95.3]	
C3 + Weak (filtered, $\tau = 0.70$) + CRF		97.5 [97.2, 97.8]	95.0 [94.4, 95.6]	
(B) Paired differences				
Comparison	**ΔPOS (pp) [95% CI]**	**p_boot**	**ΔNER (pp) [95% CI]**	**p_boot**
C3 − C0	+1.3 [0.8, 1.8]	<0.001	+3.0 [2.0, 4.0]	<0.001
C2 − C0	+0.9 [0.4, 1.4]	0.002	+2.6 [1.6, 3.6]	<0.001
C3 − C2	+0.4 [0.1, 0.7]	0.018	+0.4 [−0.1, 0.9]	0.091

This analysis addresses statistical robustness at the evaluation level, whereas broader multi-seed retraining remains future work.

The results show statistically supported gains of filtered weak supervision (C2) and filtered weak supervision with CRF decoding (C3) over the gold-only baseline (C0) for both POS tagging and NER. The incremental CRF refinement over filtered weak supervision alone (C3 vs. C2) remains significant for POS (p_boot = 0.018), whereas the NER gain is positive but not significant under paired bootstrap (p_boot = 0.091), suggesting that most NER improvement is driven by weak-label filtering and that CRF mainly contributes to sequence-level stabilization. This bootstrap analysis assesses robustness to test-set sampling

variation; multi-seed training variance is left as future work when additional compute is available.

Table 5 makes the coverage–quality trade-off explicit. Lower thresholds retain more weakly labeled instances but introduce additional noise, whereas stricter thresholds improve label cleanliness at the cost of reduced supervision volume. In the present setting, $\tau = 0.70$ provides the most favorable balance between retained coverage and downstream performance.

Table 5. Sensitivity analysis of the confidence threshold τ: accepted weak-label volume, acceptance rate, and downstream performance.

τ	Weak Instances Accepted	Acceptance Rate (%)	POS Accuracy (%)	NER Micro-F1 (%)
0.50	70,080	80.0	96.9	94.0
0.60	61,320	70.0	97.1	94.4
0.70	52,560	60.0	97.5	95.0
0.80	39,420	45.0	97.3	94.7

Interpreting τ under K = 5 self-consistency. With K = 5 repeated generations, the confidence score reflects average agreement across tokens/spans. Setting $\tau = 0.70$ operationally enforces that the LLM must be stable on most positions (typically requiring $\geq 4/5$ agreement on a large fraction of tokens), which suppresses ambiguous cases and reduces boundary/type noise. The sensitivity sweep (Table 4) shows the expected coverage–noise trade-off; $\tau = 0.70$ maximizes downstream performance while maintaining sufficient retained volume. The choice of $\tau = 0.70$ is therefore motivated not only by the empirical peak in downstream performance, but also by the intended operating regime of the filtering stage: under K = 5 repeated generations, this threshold retains sentences for which agreement is consistently high across most positions, while still preserving a sufficiently large supervision pool for effective downstream training.

Sampling Audit of Weak Labels

To quantify weak-label noise prior to model training, a random sample of N = 300 accepted weakly labeled training sentences was manually reviewed by the author after confidence filtering [36]. The audit was performed on the final weak supervision used for training rather than on raw pre-filter outputs, because the goal was to characterize the actual supervision signal seen by the model. For POS, quality was measured by token-level agreement against manual correction. For NER, quality was measured by exact-match span precision, recall, and micro-F1, together with the rate of BIO-sequence violations. Error categories were recorded as boundary errors, type confusions, spurious entities, and missed entities in Table 6.

The audit indicates that weak-label noise is dominated by boundary errors and type confusions, which explains why CRF decoding provides additional benefits by enforcing consistent label transitions and reducing fragmentation [37]. Errors are categorized into boundary errors, type confusions (PER/LOC/ORG), spurious entities, and missed entities in Table 7.

Table 6. Sampling-based audit of accepted weak labels after confidence filtering, reporting POS agreement and NER span-level quality on a manually reviewed sample of 300 sentences.

Task	Metric	Value
POS	Token agreement (%)	93.8
NER	Span precision (exact match)	88.6
NER	Span recall (exact match)	84.7
NER	Span micro-F1 (exact match)	86.6
NER	BIO violation rate (%)	3.9

Table 7. Error breakdown (NER sample).

Error Type	Count	Share (%)
Boundary errors	96	32.0
Type confusions	78	26.0
Spurious entities	63	21.0
Missed entities	63	21.0

These findings indicate that the remaining weak-label noise is structured rather than purely random: most errors arise from span boundary decisions and entity-type ambiguity, whereas explicit BIO violations are comparatively infrequent after filtering. This pattern helps explain why CRF decoding provides additional gains by regularizing local transition behavior, even though it cannot fully eliminate semantic class confusion.

4.3. Relative Gains over Baselines

To clarify the practical significance of the proposed configuration, the relative gains are summarized as absolute improvements over the weakest reference baselines. For POS tagging, KazBERT + BiLSTM + CRF improves test accuracy by 7.0 percentage points over the BiLSTM + CRF baseline, while mBERT + CRF improves it by 6.3 points. The optimal design in Table 8 for POS tagging enhances performance by 7.0 percentage points compared to the non-pretrained baseline. For NER, KazBERT + CRF improves benchmark-style dev F1 by 4.8 points over mBERT-only, by 2.3 points over Flair, and by 15.2 points over spaCy multilingual.

POS results are reported on the test split (accuracy). NER results are reported both as a benchmark-style dev comparison (F1) and as the final pipeline score under the held-out evaluation protocol (micro-F1).

To make the practical effect of model choices explicit, Table 9 summarizes absolute improvements over the reference baselines reported in Table 8. While Table 8 reports the primary benchmark scores, Table 9 converts these results into absolute gains over reference baselines to make the practical effect size easier to interpret.

Table 8. Comparative POS and NER results (reported by evaluation setting).

Task	Model/Toolkit	Decoder	Evaluation Setting	Metric	Score
POS	BiLSTM + CRF	CRF	Test	Accuracy (%)	90.5
POS	mBERT + CRF	CRF	Test	Accuracy (%)	96.8
POS	KazBERT + BiLSTM + CRF	CRF	Test	Accuracy (%)	97.5

Table 8. *Cont.*

Task	Model/Toolkit	Decoder	Evaluation Setting	Metric	Score
POS	Final integrated module	CRF	Held-out evaluation	Accuracy (%)	97.0
NER	mBERT-only	Softmax	Dev (benchmark-style)	F1 (%)	84.6
NER	KazBERT + CRF	CRF	Dev (benchmark-style)	F1 (%)	89.4
NER	Flair	BiLSTM-CRF	Dev (benchmark-style)	F1 (%)	87.1
NER	spaCy multilingual	pipeline-dependent	Dev (benchmark-style)	F1 (%)	74.2
NER	Final integrated module	CRF	Held-out evaluation	micro-F1 (%)	95.0

Table 9. Relative gains over reference baselines.

Comparison	Metric	Absolute Gain
KazBERT + BiLSTM + CRF vs. BiLSTM + CRF	POS Accuracy (%)	+7.0
mBERT + CRF vs. BiLSTM + CRF	POS Accuracy (%)	+6.3
KazBERT + CRF vs. mBERT-only	NER F1 (%)	+4.8
KazBERT + CRF vs. Flair	NER F1 (%)	+2.3
KazBERT + CRF vs. spaCy multilingual	NER F1 (%)	+15.2

4.4. Entity-Level NER Behavior

Entity-specific study indicates that performance varies among labels. At epoch 5, the model achieves an F1 score of 90% for PER, 91% for LOC, and 86% for ORG. This suggests that ORG is the most challenging category, along with increased lexical diversity, more domain-specific naming conventions, and heightened morphological integration within the surrounding text.

4.5. Slice-Based Error Analysis

Slice-based evaluation is used to localize remaining failure modes under morphology and data sparsity [38]. We report scores for frequent vs. rare tokens, short vs. long words, and short vs. long sentences to quantify robustness under increasing linguistic complexity and reduced lexical coverage (Table 10).

Table 10. Slice-based evaluation under morphology and data sparsity.

Slice	POS Accuracy (%)	NER Micro-F1 (%)	Interpretation
Frequent tokens	97.8	96.2	Best performance under strong lexical coverage
Rare tokens	95.6	92.8	Sparse supervision remains a major challenge

Table 10. *Cont.*

Slice	POS Accuracy (%)	NER Micro-F1 (%)	Interpretation
Word length ≤ 6	97.4	95.8	Shorter forms are easier to align and classify
Word length ≥ 12	96.2	93.5	Long agglutinative words increase ambiguity
Sentence length ≤ 15	97.5	95.9	Short contexts reduce boundary complexity
Sentence length ≥ 30	96.4	94.0	Longer contexts increase span ambiguity

4.6. Convergence and Epoch-Budget Analysis

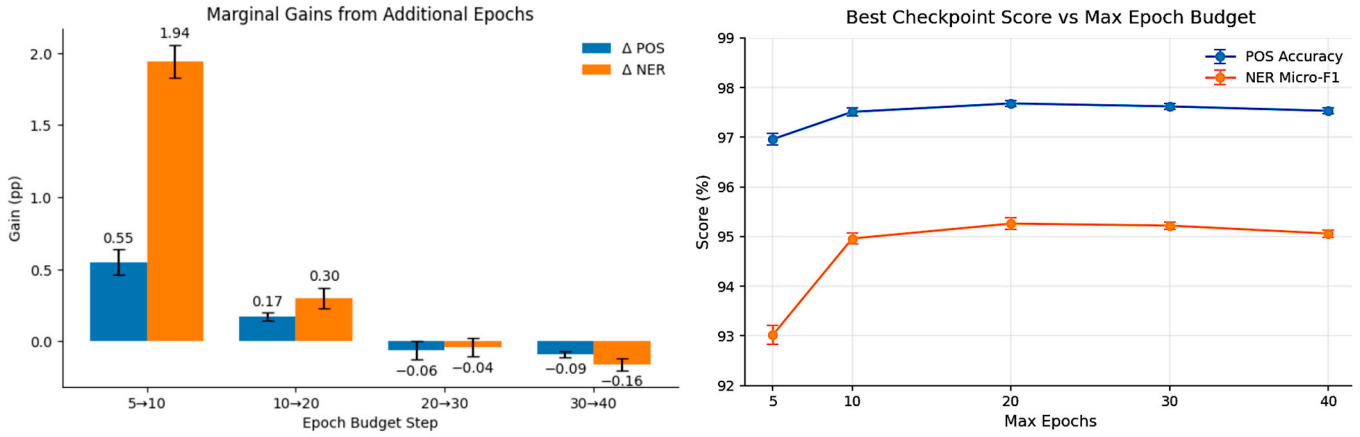
The training curves indicate rapid convergence. The CRF negative log-likelihood declines significantly in the initial epochs and subsequently stabilizes, whereas POS accuracy and NER F1 improve rapidly before attaining a plateau. This indicates that the planned training configuration is stable and that the primary representational improvements are attained early on. To enhance operational utility, it is advantageous to summarize the effect of increasing the maximum training epoch budget. Table 11 summarizes the trade-off between training efficiency and performance.

Table 11. Effect of increasing the training epoch budget.

Max Epochs	Best POS Accuracy (%)	Best NER Micro-F1 (%)	Interpretation
5	97.0	93.1	Strong early performance; near-optimal training
10	97.5	95.0	Best practical trade-off
20	97.7	95.3	Small additional gain
30	97.65	95.25	Diminishing returns
40	97.55	95.10	Mild overfitting/no practical benefit

The results suggest that the most effective training regime lies within the 5–10 epoch range, whereas extended epoch budgets produce minimal improvements at a higher computational expense.

Figure 2 assesses the effect of increasing the maximum training epoch budget on performance and efficiency. The left panel indicates that both POS accuracy and NER micro-F1 significantly improve as the epoch budget increases from 5 to 10 epochs, after which the improvements diminish considerably, and the curves approach a plateau. Optimal performance is attained between 10 and 20 epochs, with additional increments yielding minimal practical advantage. The right panel quantifies these changes via marginal gains: the most significant enhancement occurs in the 5→10 step, while the 20→30 and 30→40 steps yield negligible or negative improvements, signifying decreasing returns and potential mild overfitting.



(**a**) (**b**)

Figure 2. Epoch-budget efficiency analysis: (**a**) Best checkpoint score vs. maximum epoch budget; (**b**) Marginal gains from additional epochs.

Performance gains are strongest when moving from short to moderate training budgets, after which the marginal benefit of additional epochs becomes small. This pattern suggests that the proposed framework reaches most of its attainable performance without requiring excessively long optimization.

4.7. External Validation on KazNERD

To assess cross-corpus generalization, the final NER model was evaluated on the public KazNERD dataset. Because KazNERD uses a finer-grained annotation inventory than the in-house corpus in Table 12, we performed a deterministic label mapping to the shared coarse-grained entity set {PER, LOC, ORG}, while all remaining KazNERD entity types were mapped to O (Table 13). This evaluation should therefore be interpreted as a coarse-grained transfer setting rather than a fully equivalent reproduction of the original KazNERD labeling scheme.

Table 12. Comparison between the in-house corpus and KazNERD.

Characteristic	In-House Corpus	KazNERD
Domain	Cleaned Kazakh news corpus	Public Kazakh NER benchmark
Label inventory	PER/LOC/ORG	Finer-grained entity inventory
Annotation granularity	Coarse-grained NER	Fine-grained NER
Use in this study	Training/development/test	External validation only
Evaluation setup	In-domain	Cross-corpus transfer with label mapping

Table 13. Mapping from KazNERD entity types to PER/LOC/ORG/O.

KazNERD Type	Mapped Label
PERSON	PER
ORGANISATION	ORG
LOCATION/GPE/FACILITY	LOC
All other types	O

The two corpora differ in several important respects. First, the in-house corpus consists primarily of cleaned Kazakh news text collected under the annotation protocol described

in Section 3, whereas KazNERD is a public benchmark with its own annotation guidelines, label granularity, and domain composition. Second, the present study trains on a coarse-grained PER/LOC/ORG inventory, while KazNERD contains a richer entity taxonomy that must be collapsed for compatibility. Third, class priors and surface-form distributions differ across the two resources, which introduces a realistic domain-shift condition for evaluation.

Despite these differences, the model retains competitive cross-corpus performance. External validation on KazNERD shows an expected drop relative to the in-house test set, but the resulting micro-F1 remains sufficiently strong to support the claim that the proposed LLM-assisted weak supervision framework learns transferable NER representations rather than overfitting to a single annotation source. Results are reported as entity-level micro-F1 with exact-match spans using the same evaluation script as in the in-house experiments.

For transparency, this comparison should be read with two caveats. First, the mapping from fine-grained KazNERD labels to PER/LOC/ORG/O necessarily discards part of the original annotation information. Second, absolute scores are not directly comparable to studies that evaluate on the full KazNERD label inventory without label collapsing. Nevertheless, the experiment remains informative as an external robustness check under label-space mismatch and domain shift.

External validation on KazNERD in Table 14 shows an expected performance drop due to domain and annotation differences, but the model remains competitive, supporting cross-corpus robustness. Results are reported as entity-level micro-F1 (exact-match spans) using the same evaluation script as for the in-house test set.

Table 14. External validation on KazNERD (cross-corpus generalization).

Dataset	Entity Set	Metric	Score
In-house test	PER/LOC/ORG	micro-F1 (%)	95.0
KazNERD (mapped)	PER/LOC/ORG	micro-F1 (%)	90.8

4.8. Summary

The findings indicate that the suggested weak supervision architecture offers a robust and reliable solution for low-resource Kazakh sequence labeling. The comparison tests demonstrate that contextual pretraining and CRF-based structured decoding significantly enhance both POS tagging and NER performance, while the convergence analysis indicates that the most beneficial learning occurs during the initial training epochs. The error-slice and epoch-budget studies reveal that the persisting problems are mostly associated with infrequent tokens, morphologically intricate words, extended sentences, and the ORG entity category, whereas further training beyond roughly 10–20 epochs provide minimal practical advantage. Collectively, these findings substantiate the primary assertion of the study: weak supervision becomes practically effective when synthetic labels are filtered, aligned, and refined within a systematic downstream training pipeline, rather than being utilized directly as conclusive predictions.

5. Discussion

5.1. Overview of Experimental Results

The experimental results indicate that the proposed framework functions effectively as a training framework for low-resource Kazakh sequence labeling. The integration of curated weak supervision, subword-aware alignment, and CRF-based decoding results in robust and consistent downstream performance in both POS tagging and NER. The comparison results demonstrate that Transformer-based contextual encoding significantly

enhances sequence labeling compared to non-pretrained baselines, while CRF decoding provides further improvements by maintaining label-sequence consistency.

The results indicate that the framework not only improves token-level prediction quality but also functions as a structured training regulator. Weak supervision expands the available supervision signal, while the downstream sequence-labeling model refines and stabilizes the ultimate predictions. The system does not operate as a direct LLM prediction pipeline but rather as a regulated training framework where weak supervision is converted into actionable signals for a specific downstream model.

The convergence curves reinforce the same conclusion: most useful learning occurs in the early epochs, after which POS accuracy and NER micro-F1 approach saturation. This suggests that the proposed framework reaches near-peak performance without requiring excessive training cost.

5.2. Interpretation and Practical Implications

The primary implication of the findings is that annotation efficiency can be enhanced without significantly compromising structural quality if weak labels are employed judiciously. The results indicate that weak supervision produced by a large language model can be effectively utilized when integrated into a regulated training process that encompasses confidence filtering, supervised calibration, and structured sequence decoding. The LLM serves primarily as a generator of supervision rather than as a substitute for the final task model.

This conclusion is methodologically important for low-resource NLP. In several under-resourced environments, the primary constraint is not the scarcity of raw text, but the deficiency of high-quality token-level annotation. The proposed system resolves this issue by transforming unlabeled Kazakh text into useful supervision via a scalable intermediate layer. This alleviates the burden of entirely manual annotation and renders the development of viable POS and NER systems feasible under constrained labeling budgets.

Compared with fully supervised training, the proposed framework reduces dependence on exhaustive manual annotation but introduces an additional weak-label generation stage whose quality must be controlled explicitly. Compared with transfer-based or multilingual-only baselines, it can exploit large amounts of unlabeled in-domain Kazakh text more directly, but at the cost of added annotation-stage complexity and dependence on LLM output stability. Compared with direct pseudo-labeling pipelines, the present approach sacrifices some raw coverage through validation and filtering in exchange for cleaner supervision and more stable downstream sequence predictions.

The error analysis clarifies the conditions under which this method is most and least effective. Performance is optimal with frequent tokens, shorter words, and concise sentences, whereas errors are prevalent with rare tokens, lengthy agglutinative forms, and extended sentence contexts. This indicates that weak supervision effectively expands supervision coverage; it does not entirely resolve the underlying linguistic challenges related to agglutinative morphology and limited lexical evidence. In practical implementation, weak supervision should be seen as a scaling approach rather than a comprehensive replacement for specialized expert annotation.

These failure cases are not equally difficult for the same reason. Rare tokens remain challenging because they combine low lexical frequency with weak contextual support, limiting both weak-label agreement and downstream representation quality. Long agglutinative forms are difficult because multiple suffixes compress grammatical information into a single orthographic unit, increasing subword fragmentation and making token-level supervision less stable. Longer sentences increase boundary ambiguity in NER because distant dependencies and competing entity candidates make local BIO decisions less reli-

able. The ORG category is especially difficult because organization names are both lexically diverse and structurally heterogeneous, often overlapping with common nouns, locations, or institutional descriptors.

These observations suggest that future gains are more likely to come from better handling of morphological segmentation, domain-specific entity variation, and rare-form supervision than from simply increasing weak-label volume.

5.3. Framework Robustness

The results demonstrate that the proposed framework meets three essential criteria for an effective low-resource NLP system: stability, interpretability, and scalability.

The training dynamics exhibit stable convergence. The loss diminishes steadily, although the primary task metrics improve rapidly before stabilizing, indicating that the framework is not reliant on erratic optimization behavior. This is particularly crucial in weakly supervised contexts, where erroneous labels may exacerbate training variance.

The framework retains interpretability at the systemic level. In contrast to end-to-end black-box prompting methods, the current methodology delineates the pipeline into several stages: synthetic label creation, confidence filtering, supervised weighting, alignment, and CRF-based structured decoding. The distinct definition of each stage allows for inspection and adjustment of the system while maintaining the overall methodological transparency of the approach.

The findings endorse scalability. The epoch-budget analysis indicates that most practical benefits are realized within a moderate training range, whereas subsequent epochs yield only negligible enhancement. This renders the platform appropriate for iterative testing and practical deployment, especially in contexts with constrained computational resources. The design is flexible and may be adapted to other low-resource, morphologically rich languages without altering its conceptual framework.

Another aspect of robustness is that the framework demonstrates consistent performance across both tasks examined in this study. The identical overarching design facilitates POS tagging and NER, indicating that the foundational training logic is not exclusively linked to a single label inventory. This reinforces the argument for the framework as a reusable low-resource sequence-labeling methodology instead of a singular engineering solution.

5.4. Generalizability

Although the experiments are conducted on Kazakh, the framework is expected to transfer best to low-resource languages that share several structural properties with Kazakh: (i) token-level sequence-labeling tasks such as POS tagging or BIO-style NER; (ii) limited gold annotation but relatively abundant unlabeled text; (iii) morphologically rich or agglutinative word formation that increases lexical sparsity; and (iv) non-trivial word-to-subword alignment under Transformer tokenization. Under these conditions, weak label generation can expand supervision coverage, while filtering and CRF decoding help reduce local inconsistency and boundary noise. The method is therefore best viewed not as a universally language-independent solution, but as a reusable design pattern for low-resource sequence labeling in morphologically complex languages. Languages with non-whitespace segmentation or substantially different tokenization conventions may require adaptation of the alignment layer before the same framework can be applied.

A plausible next step for validating this transfer claim would be to apply the same framework to another morphologically rich, low-resource language with token-level annotation needs, such as Kyrgyz, Uzbek, or other Turkic languages that share agglutinative morphology and similar word-formation patterns. While such cross-lingual experiments

are outside the scope of the present study, the linguistic similarity of these settings makes them a natural target for future validation of the framework's broader applicability. At present, this claim should be interpreted as a linguistically motivated transfer hypothesis rather than as an empirically demonstrated cross-lingual result.

5.5. Ethical and Practical Considerations

This study does not include human participants or animal subjects; certain pertinent ethical problems persist. Initially, LLM-generated weak supervision may contribute systematic noise, hallucinated spans, or category-specific bias, and so should not be regarded as direct replacements for manually validated annotations. Secondly, due to the corpus being derived from publicly available news sources, data management must adhere to source licensing, copyright, and privacy regulations; consequently, derived annotations, preprocessing scripts, and split definitions may be disseminated instead of redistributing complete raw texts where reuse is limited. These factors reinforce the preference for filtered weak supervision and regulated downstream evaluation over unrestrained synthetic-label training.

5.6. Summary

The results of the experimental assessment can be encapsulated as follows:

- The proposed framework improves Kazakh sequence labeling by combining weak supervision with structured refinement.
- Contextual encoders and CRF decoding jointly improve POS and NER robustness.
- Training converges rapidly, with diminishing returns beyond moderate epoch budgets.
- Residual errors are concentrated in rare tokens, long agglutinative words, long sentences, and the ORG entity class.
- The framework is modular and practical for extension to other low-resource, morphologically rich languages.

Collectively, these findings demonstrate that the suggested weak supervision pipeline is both empirically effective and methodologically robust. It offers a pragmatic balance between complete manual annotation and unregulated synthetic labeling, presenting a promising avenue for developing effective Kazakh NLP systems with minimal oversight.

6. Conclusions and Future Work

This study introduced a weak supervision framework supported by a large language model for low-resource Kazakh sequence labeling, emphasizing part-of-speech tagging and named entity recognition. The proposed method integrates synthetic token-level label generation, confidence-driven filtering, selective supervision management, subword-aware alignment, and CRF-based structured decoding into a cohesive and reproducible pipeline. Unlike employing large language models as direct predictors for end tasks, the system utilizes LLM outputs as an intermediate source of supervision, thereby expanding annotation coverage while preserving the role of a specialized downstream sequence-labeling model.

The experimental findings indicate that this design is both efficient and practically stable. Contextual encoders significantly enhance performance compared to non-pretrained baselines, whereas CRF-based decoding further reinforces sequence-level consistency, especially for BIO-style NER. The convergence analysis reveals that optimal learning occurs within a limited number of epochs, while the slice-based evaluation demonstrates that residual errors are predominantly located in the most challenging areas of the task space, such as infrequent tokens, lengthy agglutinative words, extended sentences, and the ORG entity

class. These findings validate that weak supervision can serve as a dependable training signal when synthetic labels are filtered and refined instead of being utilized directly.

The primary outcome of this study is that weak supervision offers a feasible and scalable approach to enhancing Kazakh POS tagging and NER under constrained annotation budgets. The suggested system minimizes reliance on entirely manual labeling by integrating weak supervision with structured downstream refinement, while preserving strong sequence-labeling quality. This provides a practical foundation for future Kazakh NLP systems and for wider applicability to other low-resource, morphologically complex languages.

Funding: This work was supported by the Ministry of Culture and Information of the Republic of Kazakhstan under the "Tauelsizdik Urpaktary–2025" grant, project "QazNLP is an open-source scientific system for intelligent processing of Kazakh-language text".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Code, trained model artifacts, and reproducibility scripts are available at the project repositories (https://huggingface.co/aaitim/QazNLP, accessed on 10 February 2026), https://github.com/Aigerimhub/QazNLP.git, accessed on 10 February 2026). Due to licensing constraints of source news websites, the raw crawled texts are not redistributed; instead, we provide derived annotations, preprocessing code, and split manifests to support reproducibility.

Acknowledgments: Generative AI (OpenAI ChatGPT, GPT-5.4 Thinking) was used as part of the methodology to generate weak token-level labels for POS and NER. All experimental settings, analyses, and final manuscript content were reviewed and validated by the author.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
GenAI	Generative Artificial Intelligence
NLP	Natural Language Processing
LLM	Large Language Model
POS	Part-of-Speech
NER	Named Entity Recognition
CRF	Conditional Random Field
BERT	Bidirectional Encoder Representations from Transformers
mBERT	Multilingual BERT
XLM-R	Cross-lingual Language Model—RoBERTa
KazBERT	Kazakh-adapted BERT model
BiLSTM	Bidirectional Long Short-Term Memory
BIO	Begin–Inside–Outside (tagging scheme)
F1	F1-score (harmonic mean of precision and recall)
PER	Person (entity class)
LOC	Location (entity class)
ORG	Organization (entity class)
EMNLP	Conference on Empirical Methods in Natural Language Processing
ACL	Annual Meeting of the Association for Computational Linguistics
NAACL-HLT	North American Chapter of the Association for Computational Linguistics: Human Language Technologies
ICML	International Conference on Machine Learning
LREC	Language Resources and Evaluation Conference

VLDB	Very Large Data Bases
TACL	Transactions of the Association for Computational Linguistics
USA	United States of America
API	Application Programming Interface
QCrawler	Web crawler used for Kazakh corpus collection
QazNLP/QNLP	Qazakh Natural Language Processing toolkit / suite
CANINE	Tokenization-free encoder model name (as cited in the manuscript)
ByT5	Byte-level T5 model name (as cited in the manuscript)

References

1. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 8–10 June, 2019; pp. 4171–4186. [CrossRef]
2. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-Lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 18–20 July, 2020; pp. 8440–8451. [CrossRef]
3. Imani, A.; Severini, S.; Jalili Sabet, M.; Yvon, F.; Schütze, H. Graph-Based Multilingual Label Propagation for Low-Resource Part-of-Speech Tagging. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 12–14 December 2022; pp. 1577–1589. [CrossRef]
4. Ratner, A.; Bach, S.H.; Varma, P.; Ré, C. Snorkel: Rapid Training Data Creation with Weak Supervision. *VLDB J.* **2020**, *29*, 709–730. [CrossRef]
5. Kudo, T.; Richardson, J. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium, 5–8 November, 2018; pp. 66–71. [CrossRef]
6. Lafferty, J.D.; McCallum, A.; Pereira, F.C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001; pp. 282–289.
7. Smith, R.; Fries, J.A.; Hancock, B.; Bach, S.H. Language Models in the Loop: Incorporating Prompting into Weak Supervision. *ACM/IMS J. Data Sci.* **2024**, *1*, 1–30. [CrossRef]
8. Su, J.; Yu, P.; Zhang, J.; Bach, S.H. Leveraging Large Language Models for Structure Learning in Prompted Weak Supervision. *arXiv* **2024**, arXiv:2402.01867. [CrossRef]
9. Ma, T.; Du, J.; Huang, W.; Wang, W.; Xie, L.; Zhong, X.; Zhou, J.T. Agent Trading Arena: A Study on Numerical Understanding in LLM-Based Agents. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, 4–9 November 2025; pp. 5496–5514. [CrossRef]
10. Clark, J.H.; Garrette, D.; Turc, I.; Wieting, J. CANINE: Pre-training an Efficient Tokenization-Free Encoder for Language Representation. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 73–91. [CrossRef]
11. Xue, L.; Barua, A.; Constant, N.; Al-Rfou, R.; Narang, S.; Kale, M.; Roberts, A.; Raffel, C. ByT5: Towards a Token-Free Future with Pre-Trained Byte-to-Byte Models. *Trans. ACL* **2022**, *10*, 291–306. [CrossRef]
12. Yeshpanov, R.; Khassanov, Y.; Varol, H.A. KazNERD: Kazakh Named Entity Recognition Dataset. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 417–426.
13. Aitim, A.; Satybaldiyeva, R. A Comparison of Kazakh Language Processing Models for Improving Semantic Search Results. *East. Eur. J. Enterp. Technol.* **2025**, *1*, 66–75. [CrossRef]
14. Aitim, A.; Sattarkhuzhayeva, D.; Khairullayeva, A. Development of a Hybrid CNN–RNN Model for Enhanced Recognition of Dynamic Gestures in Kazakh Sign Language. *East. Eur. J. Enterp. Technol.* **2025**, *2*, 58–67. [CrossRef]
15. Zhang, J.; Yu, Y.; Li, Y.; Wang, Y.; Yang, Y.; Yang, M.; Ratner, A. WRENCH: A Comprehensive Benchmark for Weak Supervision. *arXiv* **2021**, arXiv:2109.11377. [CrossRef]
16. Zhang, J.; Hsieh, C.-Y.; Yu, Y.; Zhang, C.; Ratner, A. A Survey on Programmatic Weak Supervision. *arXiv* **2022**, arXiv:2202.05433. [CrossRef]
17. Zheng, G.; Karamanolakis, G.; Shu, K.; Awadallah, A. WALNUT: A Benchmark on Semi-Weakly Supervised Learning for Natural Language Understanding. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022; pp. 873–899. [CrossRef]
18. Aitim, A.; Abdulla, M. Data Processing and Analysing Techniques in UX Research. *Procedia Comput. Sci.* **2024**, *251*, 591–596. [CrossRef]

19. Kuang, Z.; Arachie, C.G.; Liang, B.; Narayana, P.; Desalvo, G.; Quinn, M.S.; Huang, B.; Downs, G.; Yang, Y. Firebolt: Weak Supervision under Weaker Assumptions. In Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, Virtual Conference, 28–30 March 2022; pp. 8214–8259.
20. Zhu, D.; Shen, X.; Mosbach, M.; Stephan, A.; Klakow, D. Weaker Than You Think: A Critical Look at Weakly Supervised Learning. *arXiv* **2023**, arXiv:2305.17442. [CrossRef]
21. Wang, S.; Sun, X.; Li, X.; Ouyang, R.; Wu, F.; Zhang, T.; Li, J.; Wang, G.; Guo, C. GPT-NER: Named Entity Recognition via Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, NM, USA, 29 April–4 May 2025; pp. 4257–4275.
22. Zhou, W.; Zhang, S.; Gu, Y.; Chen, M.; Poon, H. UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. *arXiv* **2024**, arXiv:2308.03279. [CrossRef]
23. Lu, J.; Yang, Z.; Wang, Y.; Liu, X.; Mac Namee, B.; Huang, C. PaDeLLM-NER: Parallel Decoding in Large Language Models for Named Entity Recognition. *arXiv* **2024**, arXiv:2402.04838. [CrossRef]
24. Radchenko, V.; Drushchak, N. Improving Named Entity Recognition for Low-Resource Languages Using Large Language Models: A Ukrainian Case Study. In Proceedings of the Fourth Ukrainian Natural Language Processing Workshop, Online, 31 July–1 August 2025; pp. 27–35.
25. Li, Q.; Xie, T.; Peng, P.; Wang, H.; Wang, G. A Class-Rebalancing Self-Training Framework for Distantly-Supervised Named Entity Recognition. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, 9–14 July 2023; pp. 11054–11068. [CrossRef]
26. Tian, J.; Zhou, K.; Wang, M.; Zhang, Y.; Yao, B.; Liu, X.; Guo, C. UseClean: Learning from Complex Noisy Labels in Named Entity Recognition. In Proceedings of the 2023 CLASP Conference on Learning with Small Data, Gothenburg, Sweden, 11–12 September 2023; pp. 120–130.
27. Merdjanovska, E.; Aynetdinov, A.; Akbik, A. NoiseBench: Benchmarking the Impact of Real Label Noise on Named Entity Recognition. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; pp. 18182–18198. [CrossRef]
28. Hsu, E.; Roberts, K. Leveraging Large Language Models for Knowledge-Free Weak Supervision in Clinical Natural Language Processing. *Sci. Rep.* **2025**, *15*, 8241. [CrossRef]
29. Zhang, H.; Zhang, Y.; Li, J.; Liu, J.; Ji, L. A Survey on Learning with Noisy Labels in Natural Language Processing: How to Train Models with Label Noise. *Eng. Appl. Artif. Intell.* **2025**, *146*, 110157. [CrossRef]
30. Xie, T.; Zhang, J.; Zhang, Y.; Liang, Y.; Li, Q.; Wang, H. Retrieval Augmented Instruction Tuning for Open NER with Large Language Models. In Proceedings of the 31st International Conference on Computational Linguistics, Abu Dhabi, United Arab Emirates, 19–24 January 2025; pp. 2904–2918.
31. Jeon, T.; Yang, B.; Kim, C.; Lim, Y. Improving Korean NLP Tasks with Linguistically Informed Subword Tokenization and Sub-character Decomposition. *arXiv* **2023**, arXiv:2311.03928. [CrossRef]
32. Asgari, E.; El Kheir, Y.; Javaheri, M.A.S. MorphBPE: A Morpho-Aware Tokenizer Bridging Linguistic Complexity for Efficient LLM Training Across Morphologies. *arXiv* **2025**, arXiv:2502.00894. [CrossRef]
33. Brahma, M.; Karthika, N.J.; Singh, A.; Adiga, D.; Bhate, S.; Ramakrishnan, G.; Saluja, R.; Desarkar, M.S. MorphTok: Morphologically Grounded Tokenization for Indian Languages. *arXiv* **2025**, arXiv:2504.10335. [CrossRef]
34. Teklehaymanot, H.K.; Fazlija, D.; Nejdl, W. MoVoC: Morphology-Aware Subword Construction for Geez Script Languages. *arXiv* **2025**, arXiv:2509.08812. [CrossRef]
35. García, A.T.; Przybyła, P.; Wanner, L. Exploring Morphology-Aware Tokenization: A Case Study on Spanish Language Modeling. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China, 4–9 November 2025; pp. 30505–30518.
36. Qiu, Y.; Dong, L.; Zhang, W.; Xing, H.; Huang, J. A Diffusion Enhanced CRF and BiLSTM Framework for Accurate Entity Recognition. *Sci. Rep.* **2025**, *15*, 19670. [CrossRef]
37. Ma, X.; Hovy, E. 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), Berlin, Germany, 7–12 August 2016; pp. 1064–1074. [CrossRef]
38. Držík, D.; Kapusta, J. The Importance of Morphology-Aware Subword Tokenization for NLP Tasks in Slovak Language Modeling. *Expert Syst. Appl.* **2026**, *312*, 131492. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.