Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Disease and Drug Extraction from Uzbek Medical Text with Grouped Validation and Confidence Diagnostics

Botir Elov

elov@navoiy-uni.uz

Alisher Navoi Tashkent State University of Uzbek Language and Literature

Ruhillo Alaev

National University of Uzbekistan named after Mirzo Ulugbek

Zilola Khusainova

Alisher Navoi Tashkent State University of Uzbek Language and Literature

Nizomaddin Khudayberganov

Alisher Navoi Tashkent State University of Uzbek Language and Literature

Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11844864/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Disease and Drug Extraction from Uzbek Medical Text with Grouped Validation and Confidence Diagnostics

Botir Elov[1,*], Ruhillo Alaev[2], Zilola Khusainova[1] and Nizomaddin Khudayberganov[1]

[1] Department of Computer Linguistics and Digital Technologies, Alisher Navoi Tashkent State University of Uzbek Language and Literature, Tashkent, Uzbekistan.

[2] Department of Information Security, National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, Uzbekistan.

***Corresponding author** Botir Elov, elov@navoiy-uni.uz

ORCID iDs B.E. 0000-0001-5032-6648 R.A. 0000-0003-3757-7711
Z.K. 0000-0003-4357-7515 N.K. 0000-0002-6213-3015

Abstract

We evaluated how reliably models identify complete disease and drug names in Uzbek medical text and how well their confidence scores reflect correct extraction. UzMedNER contained 712 annotated disease and drug mentions in 945 document records. Four multilingual transformer configurations were evaluated on the original test and on five development folds that kept documents sharing identical sentences together. Across three training seeds, the linear decoder achieved higher mean exact-match micro-F1 on the original test than a conditional random field with auxiliary cue supervision (60.8% versus 58.6%). On grouped development data, the latter configuration led (57.9% versus 53.8%). Temperature scaling reduced original-test expected calibration error for the linear decoder from 0.292 to 0.161 and increased it for the combined model from 0.171 to 0.210. For 76 held-out predictions from cue-bearing sentences at one seed, the three words ranked highest by integrated gradients covered 13.8% of annotated cue tokens on average, compared with 24.0% for random selection. Within this corpus, the evaluation partition changed the model ranking, and attribution magnitude was a poor guide to the annotated contextual cues.

Keywords Uzbek medical text, named entity recognition, grouped validation, confidence calibration, integrated gradients

Introduction

In the Uzbek medical corpus UzMedNER, the two words *surunkali gastrit* ('chronic gastritis') form a single annotated disease mention [1]. A tagger that labels only *gastrit* identifies the disease category but misses part of the name. Drug mentions present a related challenge. The corpus labels *klozapindan* as one drug token, including the ablative suffix *-dan* attached to *klozapin* ('clozapine'). Exact-match evaluation counts a mention as correct only when the model recovers both its annotated boundaries and its type. Such extraction can help locate passages for medication-safety research. Establishing whether a drug was implicated in an adverse event requires further evidence about the relation expressed in the text [2].

Mengliev et al. used spaCy for named entity recognition in Uzbek medical text [3]. Multilingual pretraining provides another starting point for the task, since an encoder such as XLM-RoBERTa can transfer representations learned across languages [4]. Even with the encoder fixed, two modelling choices may affect the boundaries it recovers. A linear decoder assigns a label to each token individually, whereas a conditional random field (CRF) also scores transitions between adjacent labels [5]. A model can additionally learn from expressions that the corpus marks for dosage, negation or uncertainty [1]. The cue spans are sparse, and the annotations do not link them to individual disease or drug mentions. Whether they help the model recover disease and drug names is therefore an empirical question.

The choice of evaluation data matters when comparing these models. Cross-corpus experiments in biomedical entity extraction have found that performance can fall when a system is applied to a corpus outside its training data [6]. In the evaluated UzMedNER release, the original test set contains 71 of the 712 target mentions, including 26 drug mentions. We retain that test and evaluate the 850 development documents across five held-out folds. In both designs, documents linked by sentences with identical text after case folding stay in the same partition. Each development prediction comes from a model for which the corresponding fold was held out during both training and checkpoint selection. The two evaluations assess the stability of model comparisons within a previously examined corpus.

Exact-match F1 summarises extraction across a set of mentions. For a reader inspecting one highlighted phrase, the question is how often spans assigned a given confidence have both the correct boundaries and the correct type. Temperature scaling can adjust model confidence after training [7], but calibration in structured prediction depends on the event to which the score refers [8,9]. Here, the confidence of a decoded span is computed from its token emission scores. CRF transitions contribute to selecting the span, yet do not enter this confidence calculation. We therefore examine calibration separately for each configuration. A missed mention lowers extraction recall and has no span score to include in the calibration analysis.

Token attribution raises a third question. Integrated gradients attributes a model output to input features relative to a reference input [10]. For a predicted drug mention, masking the drug name itself may substantially lower confidence even if the model makes little use of the surrounding words. We therefore compare attribution-ranked tokens with both uniformly sampled tokens and a control that selects the mention first. Following the distinction between rationale overlap and perturbation-based evaluation [11], we measure coverage of annotated cue tokens separately from the change in confidence under masking. Assessing whether an explanation helps a human reviewer would require a human evaluation [12].

We investigate three questions on the same Uzbek benchmark. First, how does the relative exact-match performance of linear and CRF decoders, with and without auxiliary cue supervision, vary between the original test and grouped development evaluation? Second, how does validation-fitted temperature scaling change the reliability of confidence scores for the spans each configuration predicts? Third, how does selecting tokens by integrated gradients compare with the two controls in coverage of annotated cues and in the change in confidence when selected tokens are masked?

Results

Corpus composition and evaluation partitions

Of the 2,926 annotated mentions in UzMedNER, 452 disease mentions and 260 drug mentions met the extraction criteria. The original test contained 71 of these 712 mentions across 140 sentences from 95 document identifiers. The five grouped development folds together contained 641 mentions across 1,170 sentences from 850 identifiers (Table 1). Seven original-test sentences and 80 development sentences contained annotated contextual cues. None of the 140 original-test sentences contained both a gold disease mention and a gold drug mention. The proposed sentence-ranking endpoint therefore had no positive cases from which to estimate recall.

Exact-match disease and drug extraction

The linear decoder (Linear) had the highest mean exact-match micro-F1 on the original test, at 60.8 ± 2.4%. The conditional random field decoder with auxiliary cue supervision (UCEE) scored 58.6 ± 2.0%. Their disease F1 scores were nearly identical (55.8% and 55.7%, respectively), while Linear achieved a higher drug F1 (70.8% versus 63.6%). On the grouped development data, UCEE ranked first at 57.9 ± 1.5% and Linear scored 53.8 ± 2.4%. Here, the larger class-specific difference was for diseases (58.6% versus 52.8%), with a smaller difference for drugs (56.8% versus 55.4%). Table 2 gives means and sample standard deviations across three seeds. For each seed, development F1 was calculated after pooling its five out-of-fold (OOF) prediction sets.

Linear scored higher than UCEE at each seed on the original test. On development OOF, UCEE led by 4.6, 5.5 and 2.2 percentage points at seeds 13, 42 and 87, respectively (Fig. 1). At seed 42, the paired UCEE-minus-Linear difference was −2.1 points on the original test (95% interval −14.9 to 7.7) and +5.5 points on development OOF (1.8 to 9.2). The intervals included zero at all three seeds on the original test and at seed 87 on development OOF (−0.9 to 5.3). These intervals reflect resampling of 91 original-test or 844 development document groups, conditional on the fitted models (Supplementary Table S3c). Averaged across cue conditions, switching from linear to CRF decoding shifted mean micro-F1 by −1.8 points on the original test and +2.5 points on development OOF. Averaged across decoders, adding cue supervision shifted it by −0.4 and +1.6 points, respectively (Supplementary Table S3d).

On the original test, the training lexicon scored 40.3% micro-F1 and a bidirectional long short-term memory model with CRF decoding scored 43.9 ± 4.1% (Supplementary Table S3b). A lexical similarity screen identified three test sentences with close matches across partitions. After their exclusion, the remaining 137 sentences retained all 71 gold mentions. Micro-F1 was unchanged for all four transformer configurations at every seed (Supplementary Table S5).

Calibration and selective prediction

Temperature scaling reduced negative log-likelihood, Brier score and expected calibration error (ECE) for both linear configurations on both evaluation sets (Table 3). For Linear, mean ECE fell from 0.292 to 0.161 on the original test and from 0.306 to 0.054 on development OOF. For UCEE, it rose from 0.171 to 0.210 and from 0.180 to 0.205, respectively. On development OOF, UCEE's negative log-likelihood decreased slightly (0.760 to 0.752) while its Brier score increased (0.270 to 0.275). The direction of the ECE change for each configuration persisted with five, ten and fifteen equal-frequency bins (Supplementary Table S4c).

The predicted spans stayed fixed under scaling, but some changed their order when ranked by confidence. On the original test at seed 42, scaling reversed the order of 142 of 2,556 pairs of Linear predictions and 149 of 2,211 pairs of UCEE predictions (Supplementary Table S4a). At the historical cutoff of 0.50, scaling reduced the number of UCEE predictions retained from 21 of 67 to eight of 67. Figure 2 shows how the fraction of incorrect predictions varied with the proportion retained. This analysis measures errors among the retained spans. A missed gold mention lowers extraction recall and has no confidence score to rank.

Attribution and coverage of annotated cues

At seed 42, all 67 original-test UCEE predictions entered the masking analysis. Nine predictions from seven documents occurred in sentences with annotated cues. For these nine, the top three words selected by integrated gradients (IG) covered 0.087 of the annotated cue tokens on average. Selecting five raised coverage to 0.193. The development analysis covered 76 UCEE predictions in cue-bearing sentences from 43 document identifiers in 42 groups (Supplementary Table S6a).

On development OOF, three words selected using 32-step IG yielded mean cue recall of 0.138, compared with 0.240 for random selection and 0.189 for the entity-first control (Table 4). The paired IG-minus-random difference was −0.102 (95% interval −0.177 to −0.033). Masking the selected words reduced mean confidence in the fixed UCEE span by 0.279 for IG, 0.089 for random selection and 0.299 for the entity-first control. The paired difference for IG was +0.190 relative to random selection (0.150 to 0.232) and −0.020 relative to entity-first selection (−0.043 to 0.001). The ordering of these mean masking effects was the same for all four configurations.

Increasing the integration resolution from 32 to 256 steps changed mean UCEE cue recall at three words from 0.138 to 0.140. Between 128 and 256 steps, the mean fraction of top-three words shared by the two rankings was 0.996. The absolute change in the confidence drop after masking those words averaged 0.005 and reached 0.356 for one prediction. At 256 steps, the median absolute difference between the summed signed attributions and the change in log confidence from the reference input was 0.0075 (Supplementary Table S7).

Keeping the selected words together with the target mention left unequal amounts of visible text. For IG and random selection, the respective means were 3.145 and 4.022 words, and the confidence drops were 0.133 and 0.136 (Supplementary Table S6e). We therefore compared how much confidence remained when the target mention was kept and each method could retain the same amount of context. With the mention alone, the mean UCEE confidence drop was 0.216. Adding three words outside the mention reduced the drop to 0.101 for 256-step IG, 0.108 for the nearest words and 0.130 for random words. With the mention included, IG and nearest-word selection retained 12.17 and 9.82 subwords on average despite identical word counts. With three outside subwords, the corresponding drops were 0.153, 0.148 and 0.162. A subword budget could retain part of a word (Supplementary Table S8).

Discussion

The choice of evaluation set changed which model appeared preferable for Uzbek disease and drug extraction. Linear led UCEE on the original test at all three seeds, and UCEE led Linear at the same seeds on the grouped development folds. The proportion of disease mentions was almost identical across evaluations (45 of 71, 63.4%, and 407 of 641, 63.5%). A shift in the overall balance between diseases and drugs therefore offers little explanation for the reversal. Other differences remain plausible. The original test included only 26 drug mentions, and each development fit used 577–579 training document identifiers compared with 756 for the original fit (Supplementary Table S2b). The design does not isolate the contributions of training size, the particular mentions encountered or their linguistic difficulty.

The grouped folds extend the comparison to more mentions while keeping documents that share identical sentences together. The development documents had been examined in earlier work, however, and the five training sets overlap. The paired intervals describe variation across held-out document groups conditional on the models fitted at each seed. Their width on the original test, and the interval that includes zero for development seed 87, limit how confidently the observed ordering can guide model selection. Cross-corpus evaluations of biomedical taggers have shown how performance changes with the source and annotation conventions [6]. An independently sampled Uzbek corpus with the same span criteria would test whether the ordering persists.

Differences between the four configurations also admit more than one interpretation. The linear decoder averages its training loss over tokens, while the CRF averages its loss over sequences and constrains label transitions. The common auxiliary-loss coefficient of 0.3 leaves the strength of cue supervision relative to the primary objective dependent on how each loss is averaged. Isolating its contribution would require matching the loss scales or varying the weight within each decoder. Prior work has improved biomedical extraction through domain-specific pretraining [13], explored weak supervision for multilingual symptom extraction [14] and evaluated an Estonian tagger on patient records [15]. The configurations in our study shared a pretrained encoder and fine-tuned it on the same corpus. This design let us compare how decoding and cue supervision interacted within UzMedNER.

Calibration mattered even when the extracted names stayed the same. On the original test, scaling improved the linear model's confidence estimates but worsened UCEE's mean error across all three calibration metrics. On development OOF, the calibration metrics for UCEE changed in different directions. The UCEE decoder uses transition scores to choose a label sequence, whereas its span confidence is derived solely from the token emission scores for that sequence. Whether including transition information improves confidence in an exactly correct span is a testable question [8,9]. Results on temperature scaling for sentence classification [16], boundary smoothing for span recognition [17] and augmentation for uncertainty estimation in NER [18] address related but distinct choices. A comparison could use the UzMedNER span criteria, fit each calibrator on data reserved separately from checkpoint selection and evaluate the same span-correctness event on a fresh source.

The masking controls clarify what a high attribution value means in this task. Removing the predicted name itself usually removes the most direct lexical evidence for its identity. That makes the entity-first control an informative comparison when assessing whether IG has located useful surrounding words. When the mention was kept intact, nearby words preserved nearly as much confidence as IG-selected words at the same three-word budget. With the budget measured in subwords, the nearest context gave a slightly smaller mean drop. Evidence that context supports a prediction is thus stronger than evidence that the gradient ranking uniquely identifies the supporting words. The rankings use attribution magnitudes, so a highly ranked word can either raise or lower the score. Even with close average agreement between integration resolutions, one mention showed a large change in masking effect. The average ranking agreement thus conceals sensitivity to integration settings for an individual prediction.

The cue annotations mark expressions such as dosage or negation anywhere in a sentence. They do not identify which expression supports a particular disease or drug mention. Selecting fewer of these cues than random selection is a finding about annotation coverage. The masking experiment asks a different question about the behaviour of a fixed model score [10,11]. Both analyses cover a single seed, and the development analysis is restricted to sentences with cues. Masking may also yield combinations of words or word fragments that a reader would rarely encounter. Determining whether an explanation helps a specialist spot an extraction error would require measuring that judgement directly [12].

The archived repair records reproduce the evaluated corpus file byte for byte (Supplementary Table S9). The available documentation leaves the original text sources, annotation adjudication and agreement, and patient or institution linkage unresolved. A formal de-identification report and an institutional ethics determination were unavailable for the evaluated release. Grouping exact duplicates and checking lexical overlap address repeated wording, while independence across clinical sources remains unknown. A next evaluation could reserve an independently annotated source, check agreement on full disease and drug boundaries, and compare both extraction and confidence on that source. A medication-safety application would additionally require expert-labelled drug–event relations and an assessment of how reviewers use the output [2]. The present evidence supports conclusions about identifying names and scoring those identifications within this Uzbek corpus.

Methods

Corpus version and target annotations

We analysed the public UzMedNER corpus [1] at the level of its original Uzbek tokens. Each input was one sentence identified by its document and sentence IDs. We used the corrected file UzMedNER.tsv from revision 625d56686a0120001d8accb517552313e5c701c3. Its SHA-256 checksum matched the archived training input. Replaying the historical row and tag corrections, document identifier reconstruction and reconciliation of six groups of duplicate annotations reproduced that file. Supplementary Tables S1 and S9 give the checksums and identify the earlier release. All analyses used the same frozen annotations and partitions.

The file contained 2,926 annotated spans in 1,310 sentences. We coded the primary task with five token labels (O, B-DISEASE, I-DISEASE, B-DRUG and I-DRUG). B marks the first token of a mention and I a continuation. We retained all 452 disease mentions and 260 drug mentions and assigned O to tokens of every other type for this task. For the auxiliary task, each token received a binary label indicating membership in a dosage, route, frequency, duration, allergy, negation or uncertainty span. These seven categories contributed 137 spans. We retained sentences without a target mention. Only the Uzbek token sequence entered the encoder. The English gloss and normalisation fields were excluded.

Grouped evaluation and model selection

We linked document identifiers whenever two documents shared an entire sentence with identical case-folded token text. A document and all other documents reachable through such links formed one group. Group membership was preserved during splitting and resampling. The original partition used split seed 42 and balanced sentence, token and annotation counts across 756 training, 94 validation and 95 test document identifiers (Table 1). As a sensitivity check, we compared sentences across the original partitions after Unicode NFKC normalisation and case folding. For sentences containing at least five distinct word tokens each, we flagged pairs whose word-set Jaccard similarity was at least 0.80. We recalculated test F1 after excluding flagged test sentences, using the existing predictions (Supplementary Table S5).

For the grouped development evaluation, we combined the original training and validation data and reserved the original test for separate reporting. We assigned the 850 development document identifiers to five outer folds, balancing document and token counts, disease and drug mention counts, and cue-token counts (split seed 20261028). For each outer fold, we divided the remaining groups between training and validation in an approximate 85:15 ratio, using the outer split seed plus the fold number. Each sentence was held out once per configuration and training seed. Within each outer training portion, we used the same inner validation set to select the checkpoint and fit the calibration temperature. Supplementary Table S2 gives the fold sizes, selected epochs and temperatures.

Encoder, decoders and auxiliary cue supervision

We fine-tuned XLM-RoBERTa-base [4] with a fast SentencePiece tokenizer, using the 768-dimensional final-layer vector at the first subword of each corpus token. The pinned encoder revision and tokenizer versions are listed in Supplementary Table S1. The longest tokenised sentence contained 279 subwords, below the configured limit of 320. A linear head produced five label scores per corpus token. The Linear configuration chose the highest-scoring label independently at each position. The first-order conditional random field (CRF) instead used learned start, transition and end scores and Viterbi decoding [5]. During training and decoding, BIO constraints penalised an initial I label and a transition to I-X unless the preceding label was B-X or I-X.

Each decoder was trained with and without a binary cue head on the same token representations. Equation (1) gives the objective for the two cue-supervised models. The primary loss was cross-entropy averaged over non-padding tokens for Linear and negative log-likelihood averaged over sentence sequences for the CRF. The cue loss was unweighted binary cross-entropy averaged over non-padding tokens. Models without the auxiliary head used the primary loss alone. UCEE denotes the CRF configuration with cue supervision.

$$L = L_{NER} + 0.3\, L_{cue} \quad (1)$$

We trained the encoder and heads with AdamW, learning rate 2×10^{-5}, batches of 16 sentences, dropout 0.1, weight decay 0.01 and gradient norm clipping at 1.0. The learning rate followed a linear schedule with 10% warm-up. Training used mixed precision on an NVIDIA RTX 3090, with CRF calculations in float32. Training ran for up to 12 epochs. The checkpoint was selected by validation exact-match micro-F1, with a minimum improvement of 0.0001 and patience of three epochs. The four configurations and seeds 13, 42 and 87 yielded 12 original fits and 60 additional fits across five folds.

Two historical baselines used only the original split (Supplementary Table S3b). A training-set lexicon stored the most frequent type for each case-folded target phrase and applied the longest available match from left to right without overlap. A one-layer bidirectional LSTM–CRF used 200-dimensional trainable word embeddings and 128 hidden units per direction. We report the lexicon once and summarise the neural baseline across the three seeds.

Predicted-span confidence and calibration

For each decoded mention *e*, we held its boundaries, type and token labels fixed. At word *i*, we divided the five emission scores by temperature T and applied the softmax to obtain the label probabilities. Equation (2) defines q_T as the geometric mean of the probabilities assigned to the decoded labels over the words in the mention. CRF transitions contribute to decoding but are absent from this confidence score. The raw score uses T equal to one.

$$q_T(e) = \exp\left[\frac{1}{|e|}\sum_{i \in e} \log p_{i,\hat{y}_i}(T)\right] \quad (2)$$

A predicted mention received a correctness indicator of one if both boundaries and type matched a gold mention in the same sentence, and zero otherwise. For each fit we estimated one scalar temperature shared by both entity types. Equation (3) averages binary negative log-likelihood over the N predicted validation mentions. We optimised log temperature in float64 with L-BFGS, starting at $T = 1$, with a strong-Wolfe line search, learning rate 0.25, up to 100 iterations and temperature bounded between 0.05 and 10. For the original Linear runs we fitted this temperature retrospectively from saved validation emissions. Each outer-fold model used its own validation-fitted temperature before its held-out predictions were pooled [7,8].

$$J(T) = -\frac{1}{N}\sum_{e=1}^{N}[c_e \log q_T(e) + (1 - c_e)\log(1 - q_T(e))] \quad (3)$$

We compared raw and scaled scores on the same predicted mentions within each model. We computed binary negative log-likelihood, the Brier score (mean squared error against the correctness indicator) and expected calibration error. For the last measure we sorted spans by confidence, formed ten nearly equal-sized bins and averaged each bin's absolute confidence–accuracy gap, weighting by its size. We repeated the calculation with five and fifteen bins. For risk–coverage curves we retained successively larger prefixes of the confidence ranking. Risk was the fraction of retained mentions that were incorrect and coverage the fraction of all predictions retained. We also counted strict pairwise order reversals after scaling and checked the historical UCEE cutoff of 0.50 on the original test at seed 42. A missed gold mention lowered extraction recall and had no predicted-span score.

Statistical analysis

The primary extraction outcome was exact boundary-and-type micro-F1 for disease and drug mentions combined. We applied the same span parser to gold and predicted labels. An I label with no open mention of the same type began a new span. We summed true-positive, false-positive and false-negative counts before calculating F1. For each seed, we pooled all five held-out development folds before computing extraction and calibration metrics. We report means and sample standard deviations over the three seed-level summaries. The original test was analysed separately. For the exploratory sentence-ranking endpoint, a positive sentence required a gold mention of both types. Recall was undefined when the test set contained no such sentence.

The primary paired contrast was UCEE minus Linear. Within each seed, we drew 5,000 bootstrap samples of document groups with replacement (seed 20261028), using exactly the same sampled groups for both models. In each draw we recomputed the difference in micro-F1 from pooled counts. The 2.5th and 97.5th percentiles provided intervals conditional on the fitted models. We summarised decoder effects by averaging paired CRF-minus-linear differences over the two cue settings and cue effects by averaging cue-minus-no-cue differences over the two decoders. We also calculated the interaction as the cue effect under CRF decoding minus the cue effect under linear decoding (Supplementary Table S3d). The overlapping outer training sets limit inference to the fitted models.

Attribution targets and perturbation controls

We used integrated gradients (IG) [10] at seed 42 to explain the log of the temperature-scaled confidence in a fixed decoded span. The original-test analysis included every predicted span. The development analysis used only held-out predictions from the 80 sentences containing at least one gold cue token. Both analyses included correct and incorrect predictions. For comparisons across models on matched predictions, we required the same document, sentence, start and end indices, and predicted type (Supplementary Table S6).

The baseline replaced ordinary subword embeddings with the padding embedding and preserved special-token embeddings, positions and attention. We integrated along the straight path in embedding space using right-endpoint sums in float32. The main analysis used 32 steps. For word ranking we summed the L2 norms of the subword attribution vectors within each corpus token, breaking ties by token index. We selected the top one, three or five corpus tokens, including punctuation, capped by sentence length. A random control used 20 uniform permutations with a reproducible seed derived from the document, sentence, span and type. The entity-first control selected mention words before nearby context, breaking distance ties by word index.

For the input sentence x and selected word set R_k, we replaced all subwords of masked words with the model's mask token and rescored the original span without decoding it again. Equation (4) measures the drop in confidence when the selected words are removed. Equation (5) measures the drop when only the selected words and the entire mention are retained. The bar over the word set denotes its complement. Both differences are on the probability scale [11].

$$Comp_k = q_T\,(e;\, x) - q_T\,(e;\, mask_{R_k}\,(x)) \quad (4)$$

$$Suff_k = q_T\,(e;\, x) - q_T\,(e;\, mask_{\overline{R_k \cup e}}\,(x)) \quad (5)$$

In equation (6), C_s contains the distinct cue-word indices anywhere in the sentence. Cue recall is the fraction of these indices selected and is undefined for a sentence without an annotated cue. We first averaged random-control values across draws within each prediction and then weighted predictions equally. For paired control differences, each resampled document group contributed the same predicted mentions to both methods. We used 5,000 group-bootstrap draws and percentile intervals.

$$CueRecall_k = \frac{|R_k \cap C_s|}{|C_s|}, \quad |C_s| > 0 \quad (6)$$

We repeated IG with 64, 128 and 256 steps on the same spans, checkpoints and temperatures (Supplementary Table S7). We compared word ranks using Spearman correlation and top-three overlap and measured the change in the confidence drop after masking those three words. For completeness, we subtracted the change in log confidence from the baseline to the input from the signed sum of attributions over embedding dimensions and subwords. The baseline forward pass used the same embedding-input path and position IDs as integration. The 256-step run provided a numerical reference for assessing stability of the lower-resolution calculations.

To control the amount of visible context, we retained every subword of the target mention and compared IG-ranked, nearest and randomly ranked context. One condition retained the mention alone, with all other nonspecial subwords masked. Other conditions added one, three or five outside words, or the same number of outside subwords, capped by availability. For a subword budget we followed word rank, retained the subwords of each word in order and stopped at the budget, allowing a partial final word. These comparisons used the fixed-span confidence drop after masking the remaining context. The subword budget constrained retention only (Supplementary Table S8).

Ethics and data provenance

We carried out a secondary analysis of a previously released text corpus [1]. This study involved no participant recruitment, intervention or collection of new patient records. The evaluated release provides limited information about the original text sources, annotator qualifications or adjudication. A formal de-identification report, documentation of permissions for the underlying texts and an institutional ethics determination for their original collection were unavailable in the study records. The grouping identifiers refer to the released documents. They cannot establish patient or institution identity.

Use of generative AI

ChatGPT/Codex (OpenAI) assisted with literature checks, preparation of analysis code, numerical consistency checks, manuscript drafting and language revision. The experimental measurements were obtained from the archived model fits, predictions and attribution runs. The authors are responsible for verifying the analyses, references and final text.

Code availability

Supplementary Data 1 includes the saved predictions and verification code used to recalculate the extraction and calibration metrics. Supplementary Software 1 provides a Python script and instructions for reproducing Figures 1 and 2 from the source data in Supplementary Data 1. Supplementary Table S1 describes the training environment and settings. The full training code and trained model checkpoints are not included in the supplementary files.

Data Availability

The UzMedNER corpus is publicly available from https://doi.org/10.57967/hf/8081. We used https://huggingface.co/datasets/uznlp-uz/uz_medner/resolve/625d56686a0120001d8accb517552313e5c701c3/UzMedNER.tsv. The dataset card specifies a CC BY 4.0 licence.

References

1. Elov, B. & Alaev, R. uz_medner. *Hugging Face* (2026). 10.57967/hf/8081
2. Wong, A., Plasek, J. M., Montecalvo, S. P. & Zhou, L. Natural language processing and its implications for the future of medication safety: a narrative review of recent advances and challenges. *Pharmacotherapy* **38,** 822–841 (2018). 10.1002/phar.2151
3. Mengliev, D. B. et al. Towards effective named entity recognition in Uzbek medical contexts. In *2024 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, 294–298 (IEEE, 2024). 10.1109/SIBIRCON63777.2024.10758445
4. Conneau, A. et al. Unsupervised cross-lingual representation learning at scale. *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451 (2020). 10.18653/v1/2020.acl-main.747
5. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. & Dyer, C. Neural architectures for named entity recognition. *Proc. NAACL-HLT*, 260–270 (2016). 10.18653/v1/N16-1030
6. Sänger, M. et al. HunFlair2 in a cross-corpus evaluation of biomedical named entity recognition and normalization tools. *Bioinformatics* **40,** btae564 (2024). 10.1093/bioinformatics/btae564
7. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. *Proc. Mach. Learn. Res.* **70,** 1321–1330 (2017). Proceedings record
8. Jagannatha, A. & Yu, H. Calibrating structured output predictors for natural language processing. *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 2078–2092 (2020). 10.18653/v1/2020.acl-main.188
9. Jiang, Z., Liu, A. & Van Durme, B. Calibrating zero-shot cross-lingual (un-)structured predictions. *Proc. EMNLP*, 2648–2674 (2022). 10.18653/v1/2022.emnlp-main.170
10. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. *Proc. Mach. Learn. Res.* **70,** 3319–3328 (2017). Proceedings record
11. DeYoung, J. et al. ERASER: a benchmark to evaluate rationalized NLP models. *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 4443–4458 (2020). 10.18653/v1/2020.acl-main.408
12. Hashemi Chaleshtori, F., Ghosal, A., Gill, A., Bambroo, P. & Marasović, A. On evaluating explanation utility for human-AI decision making in NLP. *Findings of the Association for Computational Linguistics: EMNLP*, 7456–7504 (2024). 10.18653/v1/2024.findings-emnlp.439
13. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36,** 1234–1240 (2020). 10.1093/bioinformatics/btz682
14. Sallauka, R., Arioz, U., Rojc, M. & Mlakar, I. Weakly-supervised multilingual medical NER for symptom extraction for low-resource languages. *Appl. Sci.* **15,** 5585 (2025). 10.3390/app15105585
15. Šuvalov, H. et al. Using synthetic health care data to leverage large language models for named entity recognition: development and validation study. *J. Med. Internet Res.* **27,** e66279 (2025). 10.2196/66279
16. Desai, S. & Durrett, G. Calibration of pre-trained transformers. *Proc. EMNLP*, 295–302 (2020). 10.18653/v1/2020.emnlp-main.21
17. Zhu, E. & Li, J. Boundary smoothing for named entity recognition. *Proc. 60th Annual Meeting of the Association for Computational Linguistics*, 7096–7108 (2022). 10.18653/v1/2022.acl-long.490
18. Hashimoto, W., Kamigaito, H. & Watanabe, T. Are data augmentation methods in named entity recognition applicable for uncertainty estimation? *Proc. EMNLP*, 18852–18867 (2024). 10.18653/v1/2024.emnlp-main.1049

Acknowledgements

The authors thank the colleagues who provided feedback on the annotation, experimental design and manuscript. They also thank the open-source software communities whose tools supported the analysis. Figures 1 and 2 were prepared from the study data using Matplotlib 3.10.8.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author contributions

B.E., R.A. and Z.K. conceived the study. B.E., R.A. and N.K. designed the methods and performed the formal analysis. B.E. and Z.K. curated the data. B.E. wrote the first draft. All four authors reviewed and edited the manuscript. B.E. and R.A. supervised the research. R.A., Z.K. and N.K. administered the project.

Additional information

Competing interests

The authors declare no competing interests.

Correspondence

Correspondence and requests for materials should be addressed to Botir Elov.

Supplementary information

Supplementary Information contains the corpus documentation, experimental settings, detailed results and sensitivity analyses (Supplementary Tables S1–S9). Supplementary Data 1 contains machine-readable records and verification code. Supplementary Data 2 provides the plotted values for Figures 1 and 2. Supplementary Software 1 provides the script and instructions for reproducing the two figures.

Figure legends

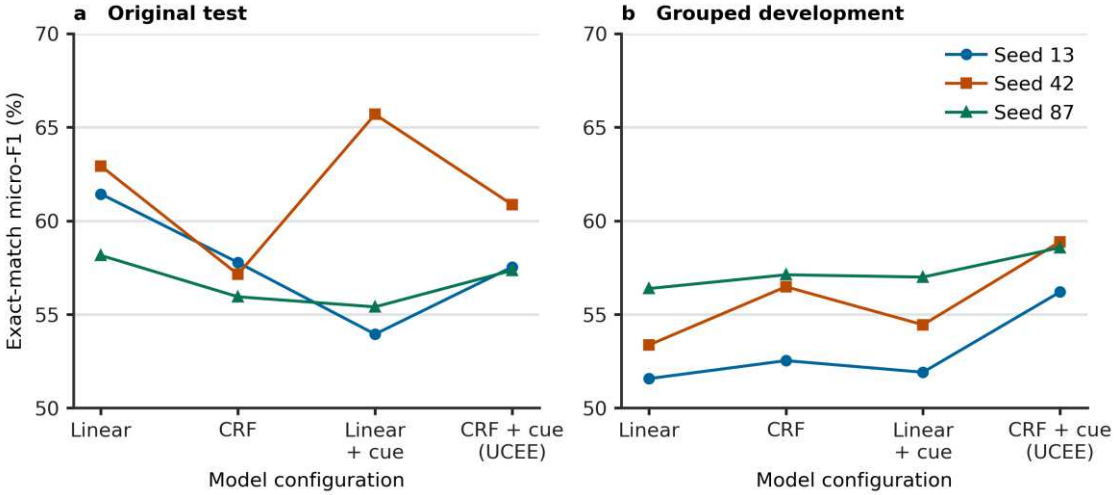

Figure 1. Exact-match micro-F1 by model configuration and seed. Panel a shows the original test. Panel b shows development out-of-fold predictions pooled within each seed. Lines connect results from the same seed across four configurations. The vertical axis gives exact span-and-type micro-F1 as a percentage. UCEE combines a conditional random field decoder with auxiliary cue supervision.

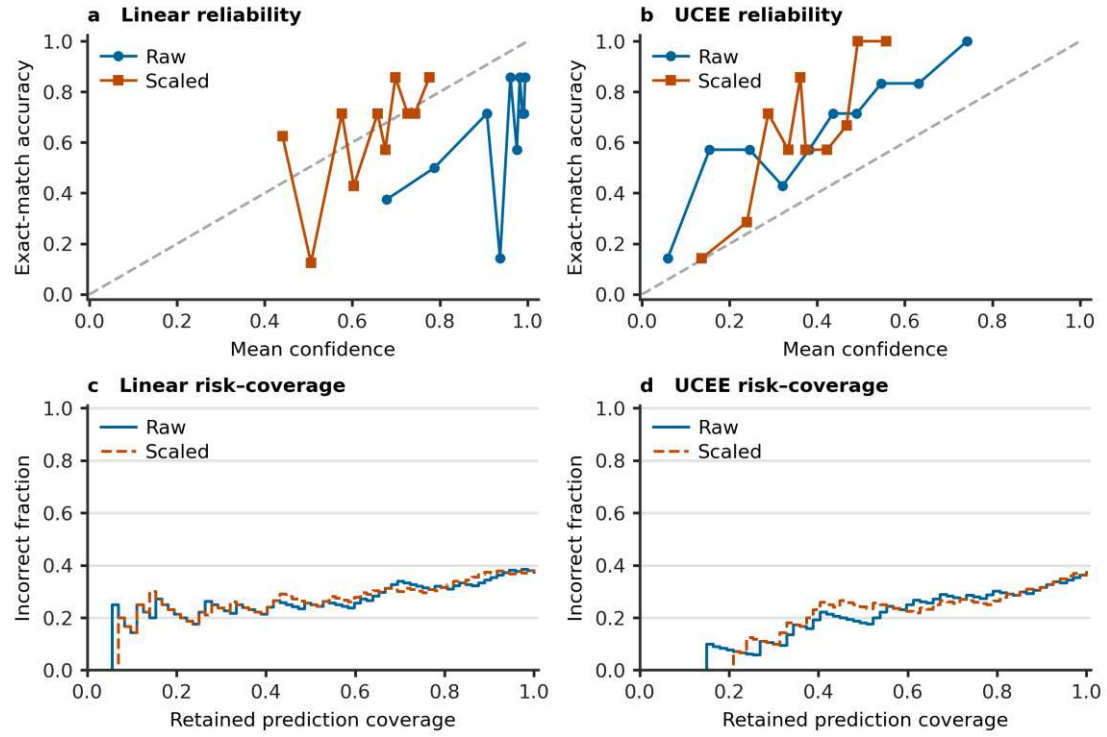

Figure 2. Reliability and risk–coverage curves on the original test at seed 42. Panels a and b compare raw and temperature-scaled confidence with exact-match accuracy for the linear decoder and the cue-supervised conditional random field (UCEE), respectively. The diagonal denotes perfect agreement. Panels c and d show the fraction of incorrect predicted entities among those retained as coverage increases. The linear decoder produced 72 predictions and UCEE produced 67.

Tables

Table 1. Corpus composition by evaluation partition.

Evaluation set	Document IDs	Sentences	Disease mentions	Drug mentions	Cue-bearing sentences
Original training	756	1,023	362	208	74
Original validation	94	147	45	26	6
Original test	95	140	45	26	7
Development OOF	850	1,170	407	234	80

Disease and drug counts are gold mentions. Cue-bearing sentences contain at least one auxiliary annotation. The development OOF row combines the original training and validation sets. The original test is held apart.

Table 2. Exact span-and-type F1 across evaluation partitions.

Configuration	Test disease	Test drug	Test micro	OOF disease	OOF drug	OOF micro
Linear	55.8 ± 3.7	70.8 ± 0.1	60.8 ± 2.4	52.8 ± 3.2	55.4 ± 1.1	53.8 ± 2.4
CRF	54.6 ± 1.9	61.9 ± 1.5	57.0 ± 0.9	54.7 ± 2.4	56.4 ± 2.9	55.4 ± 2.5
Linear + cue	53.7 ± 12.7	67.3 ± 6.1	58.3 ± 6.4	54.7 ± 2.3	54.0 ± 3.0	54.5 ± 2.5
CRF + cue (UCEE)	55.7 ± 7.7	63.6 ± 8.9	58.6 ± 2.0	58.6 ± 2.6	56.8 ± 1.9	57.9 ± 1.5

Entries are mean F1 (%) ± sample standard deviation over seeds 13, 42 and 87. OOF folds were pooled within each seed. Counts of gold target mentions are 71 for the original test and 641 for development OOF.

Table 3. Raw and temperature-scaled calibration of predicted spans.

Evaluation set and configuration	NLL raw	NLL scaled	Brier raw	Brier scaled	ECE raw	ECE scaled
Test · Linear	1.076	0.637	0.312	0.224	0.292	0.161
Test · CRF	0.666	0.685	0.233	0.245	0.174	0.215
Test · Linear + cue	1.085	0.616	0.318	0.215	0.320	0.163
Test · CRF + cue (UCEE)	0.648	0.669	0.227	0.239	0.171	0.210
OOF · Linear	0.984	0.649	0.312	0.229	0.306	0.054
OOF · CRF	0.697	0.703	0.246	0.253	0.124	0.157
OOF · Linear + cue	0.955	0.636	0.313	0.222	0.313	0.071
OOF · CRF + cue (UCEE)	0.760	0.752	0.270	0.275	0.180	0.205

Entries are means across three seeds. Lower scores indicate better performance. NLL, negative log-likelihood. ECE, expected calibration error with ten equal-frequency bins. Scores apply to predicted spans. Missed gold mentions are excluded. Supplementary Table S4 reports calibration sample sizes, confidence ordering and sensitivity to the number of bins.

Table 4. Out-of-fold attribution diagnostics at three selected tokens.

Configuration	Rationale	Predicted entities	Comp.	Suff.	Cue recall
Linear	IG	73	0.379	0.203	0.114
Linear	Random	73	0.119	0.180	0.244
Linear	Entity-first	73	0.387	0.200	0.157
CRF	IG	73	0.272	0.152	0.139
CRF	Random	73	0.084	0.137	0.243
CRF	Entity-first	73	0.299	0.148	0.191
Linear + cue	IG	72	0.372	0.151	0.122
Linear + cue	Random	72	0.123	0.161	0.242
Linear + cue	Entity-first	72	0.379	0.171	0.186
CRF + cue (UCEE)	IG	76	0.279	0.133	0.138
CRF + cue (UCEE)	Random	76	0.089	0.136	0.240
CRF + cue (UCEE)	Entity-first	76	0.299	0.149	0.189

Seed 42, cue-bearing development sentences. Comp. is the drop in fixed-entity confidence after masking the rationale. Suff. is the drop after masking context outside the union of the rationale and target mention. Smaller Suff. values correspond to greater retained confidence. Cue recall is the fraction of annotated sentence-level cue tokens selected. Random controls average 20 draws per predicted entity. Retained input sizes differ across methods (Supplementary Table S6).