

DOSTONBEK ABDURAKHMONOV

NLP / MACHINE LEARNING ENGINEER

Andijan, Uzbekistan · +998 88 338 99 01 · uzbekuzdev@gmail.com
github.com/uzbtrust · huggingface.co/uzbtrust · linkedin.com/in/dostonbek-abdurakhmonov · uzbtrust.vercel.app

PHOTO

SUMMARY

NLP engineer who builds language systems end to end — data, model, and the product around them — with a focus on low-resource Uzbek and safety-sensitive domains. Fine-tuned a byte-level ByT5 morphological analyzer to 98.9% exact match, and pretrained a 42M-parameter transformer encoder from scratch for a hybrid-retrieval chatbot. Open to full-time, part-time, on-site and remote work.

TECHNICAL SKILLS

LANGUAGES	Python · SQL · Bash
ML / DL	PyTorch · Hugging Face Transformers · scikit-learn · pandas · NumPy · TensorFlow · U-Net segmentation
NLP	Byte-level seq2seq (ByT5) · BERT pretraining (MLM, SimCSE) · BiLSTM-CRF · BPE tokenizers · TF-IDF · FAISS · RAG · NER · LLM inference and 8-bit quantization (Qwen2.5)
TOOLING	Hugging Face Hub · Streamlit · Gradio · FastAPI · Git · GitHub Actions · Linux — familiar with Docker, LangChain, LlamaIndex, Telegram Bot API, Next.js

PROJECTS

UZ-ByT5 — Explainable Uzbek Morpheme Analyzer

PyTorch · ByT5 · Transformers · Streamlit — github.com/uzbtrust/uzbek-morpheme-analyzer · huggingface.co/uzbtrust/uzbek-morpheme-byt5 · [technical report](#)

- Fine-tuned byte-level ByT5-small (**300M** parameters) to split Uzbek words into four grammatical morpheme classes and restore the dictionary form of altered stems (*o'g'lim* → *o'g'il* + *im*) — a reconstruction task that subword tokenizers cannot perform.
- Generated the training corpus instead of scraping it: applied Uzbek morphophonological rules forward over a **93K**-root lexicon and a **354**-affix taxonomy to produce **246K** examples with guaranteed-correct labels.
- Reached **98.9%** full-chain exact match and **99.4%** morpheme F1 on **3,000** held-out forms, and **100%** on the hardest subset — stems that change under affixation (**n=759**) — against **36.6%** for a rule-based baseline.
- Fixed an fp16 divergence specific to the T5 family by training in fp32, and shipped the result as a Streamlit app with a deterministic rule-based fallback, a human-correction layer, Uzbek TTS and **31** unit tests in CI.

Uzbek Telecom Operator Chatbot — Hybrid RAG + Multilingual NER

PyTorch · Transformers · FAISS · Qwen2.5-7B · Gradio — github.com/uzbtrust/uzbek-operator-rag · github.com/uzbtrust/uzbek-operator-ner

- Pretrained a **42M**-parameter BERT-style encoder from scratch — **8** layers, **512** hidden, **16K** BPE vocabulary — with masked language modelling, then fine-tuned it using SimCSE contrastive learning; **12** hours on a single Kaggle T4.
- Wrote the retrieval stack by hand — TF-IDF sparse fused with FAISS dense search, plus a confidence threshold that returns an explicit "not found" instead of letting the 8-bit Qwen2.5-7B generator hallucinate: MRR@5, NDCG@5 and Recall@5 of **0.985** across **1,015** test questions.
- Built the entity-extraction half as a BiLSTM-CRF in pure PyTorch, implementing the CRF forward algorithm and Viterbi decoder directly: F1 **0.786** on CoNLL-2003 (English) and **0.817** on WikiANN (Russian).

TriageGeist — Calibrated Emergency Triage with Uncertainty Bounds

Bio_ClinicalBERT · LightGBM · XGBoost · CatBoost · conformal prediction — github.com/uzbtrust/triagegeist · huggingface.co/spaces/uzbtrust/triagegeist

- Stacked four base models (LightGBM, XGBoost, CatBoost, PyTorch MLP) to predict Emergency Severity Index 1–5 from vitals and free-text nurse notes, combining Bio_ClinicalBERT embeddings with **~150** engineered clinical features — shock index, qSOFA, SIRS counts, ESI threshold flags, per-site normalization.
- Built for clinical safety rather than leaderboard rank: split conformal prediction for calibrated prediction sets, an asymmetric cost matrix penalising under-triage, SHAP explanations, and a fairness audit across sex, language, age and site. Deployed as a Dockerised Streamlit demo.
- The competition dataset yields near-ceiling scores (QWK **0.9998** over **80,000** patients), which is a property of the data rather than the model — the work is therefore evaluated on calibration (macro-ECE **0.00009**), conformal coverage and subgroup gaps instead.

Stenoz — Coronary Artery Stenosis Detection in X-ray Angiography

PyTorch · U-Net · scikit-image · Streamlit — github.com/uzbtrust/stenoz · huggingface.co/uzbtrust/stenoz-coronary-stenosis-unet · [technical report](#)

- Built a two-stage detector on the ARCADE dataset (MICCAI 2023, **1,500** annotated angiograms): a U-Net trained on lesion masks, paired with a geometric layer — vessel segmentation, centerline skeleton, diameter profile — that produces an auditable chart for every detection.
- Reframed the problem after finding the dataset had effectively no negative class at image level, which leaves a binary classifier nothing to learn from; localization with an interpretable explanation became the objective.
- Held-out results: **95.7%** lesion-level sensitivity and F1 **0.645** for detection, Dice **0.702** for vessel segmentation. Both networks trained from scratch on a single T4.

PUBLICATION

Abdurakhmonov, D. (2026). *From Scratch: Multilingual BiLSTM-CRF NER and Hybrid RAG Systems for an Intelligent Uzbek Telecom Operator Chatbot — Foundations for Reinforcement Learning Agents*. Preprint, Zenodo. DOI: [10.5281/zenodo.19451409](https://doi.org/10.5281/zenodo.19451409)

TEACHING

NLP & Data Science Tutor — Independent, online · 2025 – Present

- Teach Python, data science and NLP to **6–7** students, **10–12** hours per week, focused on practical model training and evaluation.

Teaching Practice, AI & Data Science — Maqsad IT Center, Qo'rg'ontepa, Andijan Region · 2025 (1 month)

- Assisted the lead instructor with lessons and assignments for an introductory AI and data science group.

COMPETITIONS & AWARDS

- **International Olympiad in Informatics (IOI)** — **73rd** nationally, Uzbekistan
- **International Olympiad in Artificial Intelligence (IOAI)** — **13th** nationally, Uzbekistan
- **Sharda University International Olympiad** — **1st** place, full four-year scholarship
- **Kaggle AIMO-3** — **749 / 3,451** teams (top **22%**)

EDUCATION

Kurgantepa Specialized School — High School Diploma · Andijan, Uzbekistan · 2024 – 2026
GPA 5.0 / 5.0 · Student of the Year, 2024–2025

FOUNDATIONS

Self-directed study behind the work above.

- **Stanford CS224N** — Natural Language Processing with Deep Learning · complete
- **Jurafsky & Martin** — *Speech and Language Processing* · complete
- **MohirDev** — Data Science and NLP course tracks · complete

LANGUAGES

Uzbek (native) · English (professional — IELTS 6.5, March 2026) · Russian (intermediate) · Arabic (beginner)