

Akam Nawzad Othman

akam.nawzad@ukh.edu.krd | akamnawzad.com | linkedin.com/in/akam-ot | github.com/akam-ot | huggingface.co/akam-ot

EDUCATION

University of Kurdistan Hewlêr (UKH)

Erbil, Iraq

BSc in Software Engineering – Graduated top 3 in programme

Oct 2021 – Jul 2025

- Thesis: *Fine-tuning XLM-RoBERTa for Named Entity Recognition in Kurmanji Kurdish*, published at WiNLP 2025, co-located with EMNLP 2025

AI Bootcamp – Certificate of Completion

Erbil, Iraq

Ovanya

Nov 2024 – Dec 2024

- Intensive programme on machine-learning fundamentals and neural networks

EXPERIENCE

AI Developer

Aug 2025 – Present

Ovanya

Erbil, Iraq

- Developed OCR and document-processing pipelines for scanned Kurdish documents
- Contributed to building one of the first OCR systems and datasets for Sorani Kurdish
- Trained and evaluated machine-learning models for text extraction from scanned pages
- Built document classification and text segmentation workflows
- Deployed model inference services with **FastAPI**
- Developed evaluation benchmarks for extraction quality and model reliability

AI Researcher

Sep 2024 – Oct 2024

UKH Artificial Intelligence and Innovation Centre

Erbil, Iraq

- Collected and cleaned a **140K+** article Kurdish news corpus for NLP research
- Built scalable web-scraping pipelines with Python and BeautifulSoup
- Contributed to developing one of the first Kurdish spell-checking datasets
- Designed annotation guidelines and supported dataset labelling workflows
- Ran model evaluations and maintained dataset versioning for reproducible research

PUBLICATIONS

KuMMLU: A Multitask Language Understanding Benchmark for Central Kurdish

2026

- Proceedings of the 10th Widening NLP Workshop (WiNLP 2026)*, co-located with EMNLP 2026, Budapest, pp. 173–180 – [paper](#) | [dataset](#)
- 3,272 multiple-choice questions across 18 subjects: 2,256 native questions from national Grade 9 and Grade 12 exams, and 1,016 translated from English MMLU
- Evaluated 32 language models; the best reach about 88% accuracy, multilingual baselines score only modestly above chance, and every model does better on translated than on native questions
- Publicly released the benchmark and evaluation code

Fine-tuning XLM-RoBERTa for Named Entity Recognition in Kurmanji Kurdish

2025

- Proceedings of the 9th Widening NLP Workshop (WiNLP 2025)*, co-located with EMNLP 2025 – [paper](#)
- Built a manually annotated Kurmanji NER dataset of 7,919 sentences (person, organisation, location)
- Fine-tuned XLM-RoBERTa to an F1 of 0.8735 (precision 0.8668, recall 0.8803); released the first public Kurmanji NER model with a web demo

PROJECTS

- Kurdish Text-to-Speech** | *PyTorch, MMS, Hugging Face* Nov 2025
- Fine-tuned Meta’s MMS TTS on a curated 10-hour Sorani dataset to produce the first open Sorani Kurdish TTS model; the demo supports Sorani and Badini
 - Improved pronunciation clarity and naturalness through audio preprocessing, transcript normalisation and evaluation
- Kurdish Spellchecker** | *ByT5, Transformers, Gradio*
- Fine-tuned ByT5-base as a sequence-to-sequence corrector for misspelled Sorani sentences; byte-level modelling handles Kurdish character variation
- KuOCR – Kurdish OCR and Document Extraction** | *Computer Vision, FastAPI* Jun 2025 – Oct 2025
- Full OCR system for Sorani Kurdish: a large real and synthetic dataset, right-to-left text rendering with bounding boxes, and a layout-aware extraction pipeline served via FastAPI
- Multilingual Legal RAG System** | *FastAPI, Vector Search* Oct 2025 – Present
- Embedding-based search and a chatbot that answers questions over official legal documents using retrieved Kurdish, Arabic and English texts, with custom text preprocessing
- Kurdish Handwritten Character Recognition** | *Python, NumPy, Streamlit* May 2025 – Jun 2025
- Implemented a 547K-parameter CNN from scratch in NumPy (manual convolution, im2col, backpropagation); 71.05% accuracy across 35 classes
- Gorrer and KurdType** | *Python, Flask, Web*
- A Latin–Arabic Kurdish script transliterator, and a curated catalogue of open-licensed fonts that support Kurdish

HONOURS

- Finalist**, 12th Undergraduate Research & Innovation Competition (URIC 2025), Abu Dhabi May 2025

TECHNICAL SKILLS

AI/ML: NLP, Deep Learning (CNNs, Transformers), Computer Vision & OCR, Speech Synthesis, LLM Evaluation, RAG

Languages: Python, C++, SQL, JavaScript

Libraries: PyTorch, TensorFlow, Hugging Face, scikit-learn, NumPy, Pandas, NLTK, Matplotlib

Backend & Tools: FastAPI, Flask, PostgreSQL, MySQL, MongoDB, Vector Databases, Docker, Git, GitHub, GitLab

Spoken: Kurdish (native), English (C1), Arabic (B1–B2)