

Murad Baghirli

+1-410-622-0676 | mbaghir1@jh.edu | [Website](#) | [GitHub](#) | [LinkedIn](#)

EDUCATION

The Johns Hopkins University

Master of Science in Engineering, Computer Science, GPA: 3.76/4.00

Baltimore, MD, USA

Aug. 2025 – May 2027

French-Azerbaijani University (UFAZ)

Bachelor of Science, Computer Science, GPA: 3.92/4.00

Baku, Azerbaijan

Sep. 2021 – Jul. 2025

EXPERIENCE

Graduate Course Assistant

Johns Hopkins University

Sep. 2026 – Present

Baltimore, MD, USA

- Support 136 students in Artificial Intelligence (EN.601.464) through weekly office hours and grading; contribute to course design, including homework and project development.

Machine Learning Engineer

National Artificial Intelligence Center (AZBA)

Jul. 2025 – Aug. 2026

Baku, Azerbaijan

- Built [E-QANUN.AI](#), the national legal search platform: 50M+ vectors across Elasticsearch and Vespa, 85% latency reduction via HNSW indexing, serving 7K+ monthly active users and 50K+ queries per month.
- Architected a multistage RAG pipeline (query normalization → spell correction → intent detection → hybrid retrieval → mGTE re-ranking → decoder generation), achieving 250ms query-to-retrieval latency.
- Designed hybrid search combining SBERT-LaBSE embeddings with lexical methods (BM25, fuzzy matching, n-gram indexing); tuned fusion weights with Optuna on synthetic queries, improving precision by 24%.
- Fine-tuned Qwen3.6-35B-A3B (MoE) with LoRA for Azerbaijani document summarization on 6×H200, adapting attention and shared-expert FFN (42M params) with frozen vision; cut eval loss from 0.561 to 0.232.

Junior Machine Learning Engineer

National Artificial Intelligence Center (AZBA)

Dec. 2023 – Jun. 2025

Baku, Azerbaijan

- Trained and evaluated SBERT/DPR/mGTE retrievers on A100/H100/H200 clusters using hard negatives, MNRL, and LambdaLoss, reaching nDCG@10 of 0.878 and MRR@10 of 0.810.
- Developed and stress-tested auto-suggestion service sustaining 500 RPS at 180ms p95 latency with zero failures.
- Designed synthetic data generation pipelines producing 1B+ tokens for retriever and reranker training.

PROJECTS

AzNeoBERT | [Project Page](#)

Jan. 2026 – Jun. 2026

- Pretrained the first modern Azerbaijani-only encoder model on a 12.2B-token corpus built from cleaned, deduplicated text sources; NeoBERT architecture with RoPE, SwiGLU, Pre-RMSNorm, FlashAttention-3.
- Built the training stack on an 8×NVIDIA H200 cluster using DeepSpeed ZeRO-2 and HuggingFace Accelerate, orchestrated with SLURM, reaching 1.24M tokens/sec with sequence packing: a 37% throughput improvement.
- Trained and benchmarked six tokenizers (SentencePiece Unigram/BPE, WordPiece; cased and uncased), selecting a 64k-vocab BPE tokenizer at 1.393 fertility vs. ModernBERT's 2.846: 51% fewer tokens on Azerbaijani.

Sentence Boundary Detector | [GitHub](#)

Nov. 2025 – Jan. 2026

- Accelerated inference with Triton Inference Server (FIL backend) on NVIDIA GPUs, enabling high-throughput processing and reducing latency from 76ms to 3ms (25x speedup).
- Trained a random forest classifier on engineered linguistic and punctuation features (0.958 F1) and wrapped it in an end-to-end pipeline handling text normalization, feature extraction, and boundary post-processing.

TECHNICAL SKILLS

Languages: Python, C, C++, Java, SQL

ML/AI: PyTorch, scikit-learn, Transformers, LLMs and VLMs, RAG, Fine-tuning, Pre-training, Recommender Systems

Training: LoRA/QLoRA, PEFT, Axolotl, Unsloth, DeepSpeed, HuggingFace Accelerate, FlashAttention, Optuna

Inference & Optimization: vLLM, NVIDIA Triton Inference Server, HuggingFace TEI, ONNX, Quantization

Search/Information Retrieval: Vespa, Elasticsearch, HNSW, BM25, Learning-to-Rank, Hybrid Search

MLOps & Infra: MLflow, Weights & Biases (W&B), Docker, Git, FastAPI, SLURM, PostgreSQL, MongoDB