

NIKITA MARKOV

Email : nikitamarkov.work@gmail.com
Blog : desire32.github.io/blog/

Phone : +357 97 533793
LinkedIn : [Nikita Markov](#)

ABOUT ME

Machine Learning Engineer with hands-on experience in ASR, NLP, and LLM fine-tuning, combining research publications, real-world projects, and strong engineering skills.

TECHNICAL SKILLS

Programming Languages: python(3.10+), golang
ML/DL: torch, transformers, sklearn, numpy, pandas
NLP/LLMs: HuggingFace, langchain, LoRA, RAG
Data/Infra: postgres, redis, kafka, docker, git, AWS EC2
Monitoring: mlflow, weights/biases, grafana

EDUCATION

University of Central Lancashire
BSc in Computer Engineering / Computing

Cyprus
Grade: First Class | 09/22 - 09/25

WORK EXPERIENCE

RIF Internship — Abasis AI

Jul 2025 – Aug 2025

Cypriot ASR dialect model, [Representation](#)

- Developed an **ASR model from scratch** in 6 weeks, based on the **Wav2Vec2** architecture.
- Integrated **KenLM language model** as an intermediate module, improving WER (word error rate) by **7%**.
- Prepared and curated training data using **pandas** and **yt-dlp**, compensating for lack of ready-made datasets.
- Trained and fine-tuned models on the **brev nvidia**.
- **Technologies:** Python, pandas, scikit-learn, Wav2Vec2, ASR, KenLM, brev-nvidia
- **Blog details:** <https://desire32.github.io/blog/pages/post-3.html>

RESEARCH PUBLICATIONS

Developing a Cyber-Physical-Social Metaverse System for Cultural Experiences

Oct 2024 - May 2025

Accepted — <https://ieeexplore.ieee.org/abstract/document/11126577/>

- Accepted into a peer-reviewed research paper with a competitive **22% acceptance rate**.
- Architected a sophisticated chatbot system using **LangChain**, integrating local LLM inference via **Ollama** with **Mistral-7B** model and persistent memory storage using **Upstash Redis**.
- Implemented a scalable **Flask-based REST API** with streaming response support and session-based chat history management, enabling seamless conversation persistence across multiple user sessions.
- Deployed the system on **AWS EC2**, demonstrating production-ready architecture with proper environment management and security considerations.
- **Technologies:** Python, **LangChain**, Flask, **Upstash Redis**, Ollama, **AWS EC2**
- **Blog details:** <https://desire32.github.io/blog/pages/post-1.html>

Enhancing Digital Heritage Experiences: Evaluating Fine-Tuned LLM Integration

May 2025 - Jul 2025

Accepted — <https://clock.uclan.ac.uk/id/eprint/56748/>

- Developed a modular fine-tuning pipeline supporting multiple architectures (TinyLlama, Mistral-7B, Llama-8B, Phi-2) with custom **qLoRA** configurations for efficient training
- Implemented a sophisticated dual-pipeline architecture combining **FAISS**-based semantic search and Word2Vec embeddings, featuring dynamic context retrieval and optimized text chunking for enhanced knowledge access
- Integrated **MLflow** for comprehensive experiment tracking, model versioning, and performance metrics visualization across different model architectures and training configurations
- **Technologies:** PyTorch, **Transformers**, **qLoRA**, **langchain-FAISS**, LangChain, **MLflow**
- **Blog details:** <https://desire32.github.io/blog/pages/post-2.html>

LANGUAGES

English (C1), Russian (Native)