# Nikita Markov

nikitamarkov.work@gmail.com | desire32.github.io/blog | github.com/Desire32

## Profile

Focused on Speech Recognition, experienced in designing and deploying end-to-end ML pipelines, fine-tuning large-scale models, and optimizing inference for edge environments.

## Skills

Python, C++, PyTorch, NumPy, Pandas, Docker, PostgreSQL, Redis, MLflow, Weights & Biases, ONNX, Git

## Experience

**RIF Internship — Abasis AI** *Jul 2025 – Aug 2025*
*Cypriot ASR dialect model*, News post

- Developed Cypriot Greek ASR system **in 6 weeks** under tight resource constraints using **Wav2Vec2** architecture, achieving performance through fine-tuning on **90,000+ audio-text pairs** and custom dataset creation from parliamentary recordings and news broadcasts.
- Implemented **KenLM n-gram language modeling** to enhance transcription accuracy, reducing Word Error Rate (**WER**) ***by 7 %*** by integrating a 6-gram model trained on 89,000+ Cypriot dialect text pairs from multiple sources.
- Built end-to-end ML pipeline including **MLflow** experiment tracking, custom data cleaning and comparative evaluation of 8+ ASR systems using ***WER/CER*** metrics.

**InSPIRE Research Center - Research Assistant** *Oct 2024 - Jan 2026*

- Led technical direction for three IEEE COMPSAC publications — selected architectures, defined experiments, and drove projects from concept to submission.

## Pet projects

- **Speech recognition pipeline** for the Cypriot dialect, managing the full lifecycle from raw audio preprocessing and phonetic transcription to MLflow experiment tracking and deployment. [GitHub]
- Hardware-Accelerated Inference Framework, a flexible CLI-based research tool for deploying quantized **LLMs on edge** hardware, supporting dynamic model swapping and vector embedding retrieval. [GitHub]
- A lightweight implementation of **LoRA (Low-Rank Adaptation)** and **RAG (Retrieval Augmented Generation)** for efficient language model fine-tuning. [GitHub]

## Education

**University of Central Lancashire** *Cyprus*
BSc in Computer Engineering / Computing *Grade: First Class | 09/22 – 06/26*

*Thesis:* ***Performance Evaluation of UE-VBS as Computational and Storage Hub (CSHs) in 6G Networks***.
Designed and implemented a prototype UE-VBS (User Equipment Virtual Base Station) communication system using **ns-3** and **Multi-Agent Reinforcement Learning (MARL)** to evaluate low-latency edge processing.

## Publications

**[IEEE COMPSAC]** Louis Nisiotis, Nikita Markov. *Quantization at the Edge: Evaluating Inference Performance and Quality for SLM Integration in Virtual Worlds —Under Review*. [Preprint]
**2.3× model size reduction** (1.1B → 470M) using **TVM** and **MLC-LLM**, Multi-scheme quantization (**INT3–INT8**), **ONNX**, **TensorRT**, (`trtexec`) on **NVIDIA Jetson**.

**[IEEE COMPSAC]** Louis Nisiotis, Nikita Markov, Charalampos Nikolaou, Aimilios Hadjiliasi, *Enhancing Digital Heritage Experiences: Evaluating Fine-Tuned LLM Integration*. [Publication]
A **qLoRA** finetuning pipeline for optimizing data access, as well as **MLflow** integration for tracking and analyzing model performance.

**[IEEE COMPSAC]** Louis Nisiotis, Charalampos Nikolaou, Nikita Markov, Aimilios Hadjiliasi. *Developing a Cyber-Physical-Social Metaverse System for Cultural Experiences* [Publication].
A chatbot system based on **LangChain** with local integration of LLM, scalable Flask-based REST API and AWS EC2.

## Languages

*English (C1), Russian (Native)*