

Mislav Jurić

mislav.juric01@gmail.com | mislavjuric.com | linkedin.com/in/mislav-juric

Skills

Domains: GenAI, LLMs, NLP, Agentic AI, RAG, Computer Vision, Traditional ML, ASR, TTS, Data Science

Programming Languages: Python, C++ , SQL

Frameworks & Libraries: PyTorch, NumPy, Hugging Face, scikit-learn, FastAPI, pandas, FlairNLP, spaCy, LlamaIndex, OpenCV, SciPy, PyInstaller, TensorFlow, Keras, Streamlit

Databases: CosmosDB, Chroma, PostgreSQL

Tools & Cloud: Git, Jupyter, Linux, Azure, Jenkins, CI/CD pipelines, GCP, Vertex AI (incl. ADK), W&B, Docker, Terraform, GNU Make, Prodigy

Experience

AI & ML Engineer and Consultant, Self-employed – Remote Oct 2023 – Present

Senior AI Engineer — Bug Triage & Root Cause Analysis Agent (Confidential Client) 2026

- Delivered a functional proof-of-concept bug triage and root cause analysis pipeline, chaining a no-code Atlassian RoVo agent into a prompt-driven GitHub Copilot agent to surface likely root causes from bug reports and shorten manual triage
- Designed an alternative system architecture in response to revised client IT policy, making development of the agent feasible
- Partnered directly with stakeholders to gather requirements and demo iterations, and received IT policy updates directly from the downstream end client

Senior AI Engineer — Enterprise Chatbot on Google Gemini (Confidential Client) 2026

- Extended a Google Gemini Enterprise chatbot with custom integrations beyond the platform's built-in connectors, enabling the agent to apply labels to emails in Gmail and move files in Google Drive
- Evaluated the platform's no-code and high-code capabilities and partnered with stakeholders to scope feasible features aligned with their needs

Senior Machine Learning Engineer — Newfire Global Partners 2025–Present

- Reduced BERT-based model inference time by roughly 50% by implementing an encoder layer-pruning technique across the production and model training frameworks
- Built an NLP model end-to-end, designing the dataset via entity prelabeling and LLM-assisted labeling, training it, and integrating it into production to power a new feature that flags abnormal records
- Retrained multiple NLP models under strict validation standards, running differential testing to verify performance before release
- Developed new features and maintained production NLP pipelines within a 7+ year-old spaCy codebase, using an in-house FlairNLP-based framework for model training, and updated Jenkins CI/CD

AI Lead — Computer Vision Advertising Project (Confidential Client) 2025

- Defined the technical architecture and established realistic delivery timelines with the CEO, tech lead, and infrastructure lead
- Identified that the existing edge hardware could not run the AI pipeline, prompting a hardware change before development committed to an unworkable target
- Built the project's first complete ML processing pipeline, running all features end-to-end where the previous prototype could not, and establishing a working baseline the incoming AI Lead reused
- Mentored and managed a university-level engineer, assigning work, providing technical direction and code review

Machine Learning Engineer — AI People, LLM-Powered Game (GoodAI) 2023–2025

- Engineered prompt-protection mechanisms that increased time-to-extract the system prompt from under 1 minute to roughly 30 minutes for technical testers, and to days for non-technical QA
- Contributed to a QLoRA training framework, adding OpenAI fine-tuning support, then inherited and used it to fine-tune local 7B LLMs (Llama, Mistral) after the lead departed
- Ran and evaluated local LLM and ASR inference using llama.cpp, whisper.cpp, and NVIDIA In-Game Inferencing (NVIGI), identifying an SDK bug that NVIDIA later patched
- Integrated local and cloud ASR/TTS, implementing a self-contained FastAPI + PyInstaller deployment for a local TTS model released under a week prior, and generated arbitrary synthetic voices by manipulating voice embeddings in vector space
- Built content-filtering and safety mechanisms plus credit-calculation logic reconciling in-game credits against actual LLM, ASR, and TTS spend, surfacing a billing discrepancy

Machine Learning Engineer, TIS Group – Remote Oct 2020 – Sept 2023

- **SENDD (Healthcare)**: Reduced pose-estimation model training time from weeks to days by moving training onto GPU, and later multi-GPU, using NVIDIA's trt_pose
- Improved pose-estimation accuracy through transfer learning and built a heatmap-feature classifier using kNN and SVM with PCA for dimensionality reduction
- **Vineyard Angel**: Led development of a drone-imagery pipeline from design to deployment, segmenting vineyard rows and gaps and estimating vine-gap counts with PyTorch Segmentation Models and OpenCV
- Automated the pipeline on Azure via systemd-orchestrated VMs, and presented the project with the agricultural partner at AI2FUTURE 2022

Software Engineering Intern, Rimac Automobili – Sveta Nedelja, Croatia July 2019 – Feb 2020

- Built C++ infrastructure around ML models within an autonomous-driving team, including systems parsing low-level radar message bytes
- Worked with Linux, Git, and NVIDIA autonomous-driving hardware under the mentorship of a senior engineer

Education

University of Zagreb, Faculty of Electrical Engineering and Computing – MS in Computer Science July 2020

University of Zagreb, Faculty of Organization and Informatics – BS in Informatics July 2018