

Uroš Milutinović

AI Engineer & Software Developer

umilutinovic25@gmail.com | +381 64 991 0200 | Belgrade, Serbia

github.com/umilutinovic25-hash | linkedin.com/in/uros-milutinovic04 | umilutinovic25-hash.github.io/portfolio | huggingface.co/uros69

PROFESSIONAL SUMMARY

AI and software engineer who ships production systems, not just prototypes. Built a RAG pipeline (LangChain, GPT-4, Pinecone) with a FastAPI + React.js service layer handling 500+ queries/day in production; fine-tuned open-source LLMs (Mistral 7B, Llama 3) with LoRA, evaluated through RAGAS and BERTScore pipelines; designed multi-agent and Claude-powered agentic systems, from a voice-controlled Revit BIM plugin to an autonomous job-application agent. Focused on production LLM systems: evaluation pipelines, monitoring, Dockerized deployment, and cost/latency tracking. B.Sc. in New Computer Technologies (Feb 2026), AWS Machine Learning Specialty certified.

KEY SKILLS

- AI / ML:** Python, PyTorch, LangChain, RAG, LLM fine-tuning (LoRA), Hugging Face, OpenAI API, Anthropic Claude API, Pinecone, multi-agent orchestration, evaluation (RAGAS, BERTScore)
- Evaluation & Observability:** RAGAS, BERTScore, llm-eval (cost and latency tracking, regression gating), structured logging, vector search (Pinecone, embeddings, sentence-transformers)
- Full-Stack:** FastAPI, React.js, Java, Spring Boot, PHP, MySQL, PostgreSQL, REST APIs
- DevOps / Cloud:** Docker, Git, Linux, n8n (workflow automation), AWS (ML Specialty certified), Agile / Scrum

WORK EXPERIENCE

AI & Software Engineer

Aug 2026 – Oct 2026

Wild Vision · Belgrade

- Built an AI-powered video post-production pipeline: Whisper transcription, YuNet face detection for automatic 1:1 reframing, Demucs voice/music separation, and burned-in captions and motion titles, turning long-form footage into social-ready cuts.
- Developed a full-stack internal HR review tool (React, Vite, Node.js, SQLite) with role-based access for HR, managers and employees, a bilingual EN/SR interface and 100+ automated tests, delivered as a working prototype.
- Added accessibility checks (axe) and mobile-friendly tables, and prepared a seeded demo environment for stakeholder review.

AI Junior Engineer

Mar 2026 – Aug 2026

Netconomy · Belgrade

- Built, tested, and refined AI/LLM-based applications and workflows as part of a production-oriented engineering team.
- Collaborated cross-functionally to integrate AI technologies into the company's software systems, working within an Agile development process.

AI & Full-Stack Engineer — Contract / Freelance

Dec 2024 – Present

Self-employed · Remote (international clients) · alongside B.Sc. studies

- Designed and shipped a RAG system for document Q&A (LangChain, GPT-4, Pinecone) running at 500+ queries per day, with a FastAPI backend for asynchronous inference and a React.js frontend with response streaming.
- Fine-tuned open-source LLMs (Mistral 7B, Llama 3) on domain-specific data using Hugging Face Transformers and LoRA adapters; added RAGAS / BERTScore evaluation and Docker deployment with structured logging.
- Delivered live client websites, including N Detailing Studio and FitPro Agency — a lead-generation SPA with EmailJS, Google Sheets CRM, Meta Pixel tracking, and an admin panel.
- Currently building a Claude-powered job-analysis agent and an LLM evaluation framework (see Featured Projects).

IT Support Engineer

Jun 2024 – Dec 2024

Microsec d.o.o. · Belgrade

- Resolved 20+ tickets per week and wrote internal runbooks that cut repeat incidents by 30%.
- Administered MySQL databases with zero data loss; configured developer workstations and supported deployment pipelines and Postman API testing.

ADDITIONAL EXPERIENCE

Truck Dispatcher & Track-and-Trace Specialist — Sky Freight Dispatch (U.S. market), Oct 2025 – May 2026 (part-time).

Coordinated U.S. freight transport and GPS tracking; managed all client and driver communication in English.

Guest Services & Operations — Work and Travel USA Program, United States, Summer 2025. Full-time guest-facing role in English; resolved a high volume of guest requests in a fast-paced international team.

FEATURED PROJECTS

Revit AI Assistant — Voice-Controlled Multi-Agent BIM Automation Plugin

AI plugin for Autodesk Revit that automates engineers' workflows directly from the 3D building model via natural-language voice commands. Multi-agent architecture: an orchestrator agent interprets intent and delegates to specialized sub-agents for quantity takeoffs, material estimation, and model manipulation.

Stack: C# / .NET, Revit API, multi-agent LLM orchestration, Speech-to-Text

JARVIS — Voice-Controlled Personal AI Desktop Assistant (Electron)

Desktop assistant driven by voice (wake word + hotkey) that answers in Serbian and English, reads the user's screen on request, and visualizes a "second brain": a 3D graph of Obsidian and memory notes, searched with local multilingual embeddings (transformers.js). Local Whisper speech recognition, Claude as the reasoning layer, a Next.js + Supabase dashboard for calendar, mail and agent status, and 190+ automated tests.

Stack: Electron, TypeScript, Node.js, Whisper, openWakeWord, transformers.js, Next.js, Supabase, Claude API

DocuMind — Private Document Q&A with Verified Citations · github.com/umilutinovic25-hash/documind

Local-first hybrid RAG Q&A with streamed, source-cited answers; evaluation caught an overstated hit@k (1.00 vs 0.73) and guided a chunking fix to 0.87.

Stack: Python, FastAPI, React, Ollama (bge-m3, gemma3), Qdrant, pytest (44 offline tests)

RAG Eval Lab — Measured, One-Change-at-a-Time RAG Improvement · github.com/umilutinovic25-hash/rag-eval-lab

Took a deliberately weak RAG system from 6/12 to 12/12 through single-change experiments, each with a spec and frozen test questions committed first; 489/508 (96%) across 22 question sets incl. blind sets. Added version-aware retrieval (9/20 → 20/20) and injection guards (6/10 → 10/10, blind); measured and rejected hybrid search and query rewriting.

Stack: Python, Ollama (bge-m3, Qwen2.5, Gemma 3), LLM-as-judge, React, Playwright, pytest (288 tests)

AppLens — Cross-Platform Application Security Analyzer · github.com/umilutinovic25-hash/applens

Built a static analyzer for macOS, iOS, Electron, Android and browser extensions, with SARIF output, release diffs and CI gating; added optional local Ollama explanations without sending reports to a hosted AI API.

Stack: Python, static analysis, reverse engineering, SARIF, Ollama

carservice-api — Car Service Workshop REST API (Java/Spring Boot) · github.com/umilutinovic25-hash/carservice-api

Spring Boot API for customers, vehicles and work orders, with entity-level status rules, price-snapshotted items and invoice generation. Flyway migrations and 30 tests, including full lifecycle integration.

Stack: Java 17, Spring Boot 3, Spring Data JPA/Hibernate, MySQL, Flyway, Maven, JUnit 5, Mockito, Docker

JobScout — AI Job-Application Agent (Claude API) · github.com/umilutinovic25-hash/jobscout

Autonomous agent that analyzes job postings against a CV: scrapes the posting, extracts structured requirements with Claude, scores the match on a calibrated 0–100 rubric with gap analysis, drafts a tailored cover letter, and tracks every application — including per-application API cost — in SQLite.

Stack: Python, Anthropic Claude API, structured outputs, httpx + BeautifulSoup, SQLite, pytest

MailGuard — Email Phishing & Fraud Scanner · github.com/umilutinovic25-hash/mailguard

Security CLI that flags whether inbox emails are safe to open: layered heuristics over full headers — SPF/DKIM/DMARC, brand-spoofing and lookalike domains, masked links, dangerous attachments — plus an LLM second-pass verdict on borderline cases; runs as a scheduled daily scan.

Stack: Python, IMAP, email authentication (SPF/DKIM/DMARC), LLM triage, launched

Credit Card Fraud Detection — Imbalanced ML + Real-Time Scoring API · github.com/umilutinovic25-hash/credit-card-fraud-detection

Fraud detection on 284k imbalanced transactions (0.17% fraud): KNN, XGBoost and LSTM ensemble with SMOTEENN achieved F1 0.82. FastAPI scoring service measured 3.5 ms p50 latency; Gradio demo simulates bank transactions.

Stack: Python, scikit-learn, XGBoost, PyTorch, imbalanced-learn, FastAPI, Gradio, pytest

dine-agent — AI Phone Host for Restaurant Reservations · github.com/umilutinovic25-hash/dine-agent

Claude-powered voice agent that answers calls and manages the booking book. Every write goes through a policy-owning engine — service windows, table capacities, dining durations, conflict detection — so the model proposes but can never invent a table.

Stack: Python, Anthropic Claude API (tool use), FastAPI, Web Speech API, pytest

llm-eval — LLM Evaluation Framework · github.com/umilutinovic25-hash/llm-eval

Framework for comparing LLMs (Claude, local models via Ollama) on the same task suite — accuracy, cost, and latency side by side. YAML-defined evals, async runner with per-request token and cost tracking, JSON snapshots for regression detection, non-zero exit codes for CI gating, and a self-contained HTML report dashboard.

Stack: Python, asyncio, Anthropic API, Ollama, YAML, pytest

Other projects · [all on github.com/umilutinovic25-hash](https://github.com/umilutinovic25-hash)

Semantic Search over 50k ArXiv papers (sentence-transformers) · Editor Workflow Assistant (Premiere Pro, Whisper) · poslovi-mcp (job-board search MCP) · Serbian LLM Benchmark (8 models) · Razumem Nalaz (on-device OCR PWA) · SafeCircle (FastAPI, Twilio) · N Detailing Studio (live site) · Car Service Platform (B.Sc. thesis)

CERTIFICATIONS

- AWS Certified Machine Learning – Specialty — AWS, 2026
- Claude Code, AI Fluency, MCP & Claude API — Anthropic Academy; Monitoring Production Agents — LangChain Academy, 2026
- AI Agents (GAIA: 100%), MCP, LLM/NLP & Deep RL — Hugging Face, 2026
- GitHub Foundations, HackerRank SQL (Advanced) & Problem Solving (Intermediate) — 2026
- Feature Engineering, Time Series & Intro to ML — Kaggle Learn; Databricks Fundamentals — Databricks, 2026
- Machine Learning & Advanced Learning Algorithms — DeepLearning.AI, 2026

LANGUAGES

- English — C1 (fluent business English)
- Serbian — native
- German — A2

EDUCATION

B.Sc., New Computer Technologies **Graduated Feb 2026**

VIŠER — College of Electrical Engineering and Computer Science, Belgrade (2022 – 2026)

Mechatronics Technician

2018 – 2022

Technical School “17. Septembar”, Lajkovac