

MAKSYM BEIEV

Applied AI Engineer · AI & Automation Lead

AI agents & automation · model fine-tuning & inference · evaluation & quality · 10+ years in video production

Fully remote · B2B or employment | max.beyev@gmail.com | +48 502 645 938 | rzmrn.com | LinkedIn | GitHub

English (professional) · Polish (native-level) · Ukrainian · Russian

I build AI systems that do real work: agent workflows and pipelines that run production, models tuned and served for the job, and the tests and metrics that keep the output reliable. In 2026 I led AI and automation at the in-house AI studio of a creator-marketing agency, now serving external clients: agent swarms run about three-quarters of production under human quality gates, self-hosted open-weight models cut compute cost per video about 6× versus a hosted API, and controlled A/B tests with custom metrics took a critical defect from 89% of frames to 0%. I work AI-native: I design the system, write the specs and tests, and direct coding agents that write most of the code. Before AI, 10+ years in video production and team leadership — so I know what good output looks like and what a client will accept.

AI agents Automation LLMs RAG MCP Fine-tuning (LoRA/DoRA) Model inference Evals Computer vision
Generative video Python Docker Video production

EXPERIENCE

AI & Automation Lead In-house AI studio of a creator-marketing agency (confidential) Apr 2026 – Present · Remote

- ◆ **Agent systems that run production** — Designed agent swarms that run about three-quarters of production stages — from pulling client references to GPU rendering, packaging and delivery — under hard budget caps, while people keep three gates: budget, identity sign-off and the final verdict. Agent-run output grew from 44% to 81% in one month; a new client's 100-clip order became 121 approved videos 19.5 hours after kickoff.
- ◆ **End-to-end AI pipelines** — Built the studio's generation pipeline: LLMs write model-specific prompts, 8+ image and video models plus pose and face models do the work, a human review gate approves and delivery is automated. 4,000+ images and videos delivered in the first four months; AI content became the agency's main traffic source.
- ◆ **Training and tuning models** — Built a LoRA/DoRA fine-tuning pipeline for character models (dataset snapshots, LLM captioning, checkpoint registration); ran open-weight video models of up to 33B parameters on 32 GB GPUs through int8/ NVFP4 quantisation and workload-aware memory envelopes; chose models and settings by benchmarks and A/B tests, not by feel.
- ◆ **Quality you can measure** — Designed metrics and test series that surface edge cases: face-embedding identity drift, pose-keypoint accuracy, pixel-level parity checks. A controlled A/B investigation took a critical identity defect from 89% of frames to 0%; fixing the brief lifted the good-clip rate on a new engine from 36% to 69%.
- ◆ **Performance and cost** — Moved video generation from paid APIs to self-hosted models on a fleet of up to 20–26 cloud GPUs: ~6× lower compute cost per video and ~89% API spend within a month. Cut pose preprocessing from 12 minutes on CPU to 45 seconds on GPU by tracing an ONNX Runtime CPU/GPU package collision; shipped a 33–36% render speed-up only after a pre-merge GPU regression test proved pixel-identical output (SSIM 1.0).
- ◆ **Distributed job execution** — Built a durable job queue (leases, heartbeats, fencing tokens, retries, GPU resource classes) and a single-ledger dispatcher that replaced failing batch jobs: 230 clips rendered overnight on up to 12 RTX 5090s; a 380-clip wave on up to 20 cloud GPUs with zero queue failures and under 1% idle time.
- ◆ **Reliable production, safe releases** — 951 videos in about 30 hours with zero render failures. Every release passes a CI gate of 1,145 automated tests, with one-command rollback and a gate that blocks deploys during live renders.
- ◆ **Team and business** — Variable cost per unit fell 45% month over month while output grew 18%. Hired and led 2–4 operators (output per operator-hour ×3.6 in one month); built pricing, packaging and the sales deck; the first external client orders followed.

Claude Code Codex MCP ComfyUI LoRA / DoRA int8 / NVFP4 ONNX Runtime Python FastAPI
Docker / CUDA GitHub Actions GPU cloud (RunPod)

Head of Content Production Kemp Center / kursyszkolenia.online (e-learning) Dec 2024 – Mar 2026 · Poznań

- ◆ **4.5× faster production** — Replaced a linear one-editor-per-lecture flow with a parallel conveyor: finishing time fell from 3 hours to 40 minutes per lecture.

- ◆ **~600 lectures across 20 courses** — Built and led a 4-person production team from scratch in a 20-person company, shipping 3–4 lectures a day.
- ◆ **AI and automation in production** — Generative backgrounds (Midjourney), animated B-roll (Runway, Kling) and AI presenters with synthetic voice (HeyGen, ElevenLabs); ExtendScript batch automation across hundreds of compositions; 221 archived lectures migrated through a five-phase automated pipeline with zero data loss.

Claude / GPT API Midjourney Runway Kling HeyGen ElevenLabs ExtendScript

Freelance Video Producer & Motion Designer Fiverr · Top Rated Oct 2022 – Dec 2024 · Remote

- ◆ **298 projects, 120 US clients** — Grew a one-person studio into a small micro-agency with subcontractors: briefs, concepts, filming, editing, grading, VFX and multi-platform delivery; zero revisions on 70% of deliveries.

Director / Lead Video Producer State Folk Ensemble “Mazowsze” — Ministry of Culture 2021 – 2022 · Otrębusy

- ◆ Full-cycle production for Poland’s national folk ensemble (170+ performers): 200+ concerts filmed in Poland and on international tours, live multi-camera broadcasts, TV spots and animated ads for Warsaw Central Station and city transit, presidential and diplomatic events.

Video Producer & Cinematographer Rozmarin Studio / Automax Detailing 2015 – 2021 · Warsaw & Kyiv

- ◆ Six years of hands-on production across automotive, advertising and creative work; grew from camera operator to studio lead.

SELECTED PROJECTS

Max Chronicle open source · MIT

Local-first memory for AI agents: MCP server, CLI and SQLite event store. Hybrid retrieval (full-text, local vector embeddings, temporal) with its own eval harness for recall quality; 700+ tests in CI. Runs my agent workflows daily.

Python MCP Retrieval Evals

LocalFlow open source · MIT

On-device speech and LLM pipeline for macOS: dictation and meeting memory with local models only, speaker separation and a searchable, cited archive; scripted model evaluation. 97 tests, CI on macOS.

Swift MLX Local LLMs Speech

RZMRN Digest autonomous LLM pipeline

15-stage LLM pipeline over 84 sources: collection, deduplication, parallel analysis, anomaly detection, translation and publishing twice a day, with a watchdog and alerting. Ran unattended for months.

Python LLM pipelines Cloudflare

SKILLS

AI agents & automation	Claude Code and Codex, multi-agent workflows with budget and safety limits, MCP servers, tool calling, structured outputs, prompt engineering, RAG, agent memory (mem0, Kuzu, local embeddings)
Models & training	LoRA/DoRA fine-tuning (ai-toolkit), open-weight model inference, quantisation (int8, NVFP4, fp8), memory and attention-kernel tuning, diffusion image and video (WAN, LTX, MiniMax H3, FLUX, Z-Image), computer vision (ViTPose, RTMPose, SAM, face embeddings), LLM and VLM APIs (Claude, GPT, DeepSeek, Qwen-VL), speech (Whisper, pyannote)
Evaluation & quality	Metric design, test series and edge cases, single-factor A/B testing, regression and parity gates, eval harnesses, human-in-the-loop review
Infrastructure & tools	Python, FastAPI, SQLite (WAL), durable job queues, Docker (CUDA images), GitHub Actions CI/CD, GPU fleet on RunPod, ComfyUI, Linux, Next.js
Leadership & delivery	Lead of a production AI platform, specs and acceptance criteria, client delivery, unit economics and forecasting, hiring and leading teams
Video production (10+ years)	Premiere Pro, After Effects, DaVinci Resolve, Cinema 4D, colour grading, motion design, live multi-camera, drone (EASA A1/A2/A3)

EDUCATION & CERTIFICATES

Information Management studies PJATK (Polish-Japanese Academy of Information Technology) 2016 – 2019

Coursework: database architecture, API design, business process engineering. Certificates: Cinema 4D (Isaevworkshop), UAV drone pilot A1/A2/A3 (EASA), colour grading (Hohlow & Sobatowski).