

MAKSYM BEIEV

Applied AI Engineer · AI & Automation Lead

Agenci AI i automatyzacja procesów · fine-tuning i inferencja modeli · ewaluacja i jakość · 10+ lat w produkcji wideo

Praca w pełni zdalna · B2B lub umowa o pracę | max.beyev@gmail.com | +48 502 645 938 | [rzmrn.com](#) | [LinkedIn](#) | [GitHub](#)
Polski (biegły) · angielski (zaawansowany) · ukraiński · rosyjski

Tworzę systemy AI, które działają na produkcji: agentów i pipeline'y automatyzujące pracę, modele dostrojone i wdrożone do konkretnych zadań oraz testy i metryki kontrolujące jakość. Od kwietnia 2026 r. odpowiadam za AI i automatyzację w wewnętrznym studiu AI agencji creator marketingu, które obsługuje dziś także klientów zewnętrznych. Agenci AI realizują tam ok. trzech czwartych produkcji, a ludzie kontrolują kluczowe etapy; modele open-weight uruchomione na własnej infrastrukturze obniżyły koszt obliczeń na film ok. 6-krotnie względem płatnego API, a kontrolowane testy A/B z autorskimi metrykami obniżyły odsetek klątek z krytyczną wadą z 89% do 0%. Pracuję w modelu AI-native: projektuję system, piszę specyfikacje i testy, a większość kodu powstaje z pomocą agentów kodujących. Wcześniej przez ponad 10 lat produkowałem wideo i kierowałem zespołami — dlatego wiem, jaki efekt klient zaakceptuje.

Agenci AI Automatykacja procesów LLM RAG MCP Fine-tuning (LoRA/DoRA) Inferencja modeli Ewaluacja modeli
Computer vision Generatywne wideo Python Docker Produkcja wideo

DOŚWIADCZENIE

AI & Automation Lead Wewnętrzne studio AI agencji creator marketingu (NDA) 04.2026 – obecnie · zdalnie

- Systemy agentowe w produkcji** — Zaprojektowałem system wieloagentowy, który realizuje ok. trzech czwartych etapów produkcji — od pobrania materiałów klienta, przez rendering na GPU, po spakowanie i przekazanie gotowych plików — w ramach sztywnego budżetu. Człowiek decyduje w trzech punktach kontrolnych: budżet, zgodność postaci i ostateczna akceptacja. Udział produkcji realizowanej przez agentów wzrósł w miesiąc z 44% do 81%; nowy klient zamówił 100 klipów, a po 19,5 godziny miał już 121 zaakceptowanych filmów.
- Pipeline'y AI end-to-end** — Zbudowałem pipeline generacyjny studia: LLM tworzą prompty dopasowane do każdego modelu, generację obsługuje ponad 8 modeli obrazu i wideo oraz modele pozy i twarzy, człowiek zatwierdza wynik, a przekazanie materiałów jest zautomatyzowane. W pierwszych czterech miesiącach — ponad 4000 zdjęć i filmów; treści AI stały się głównym źródłem ruchu agencji.
- Trening i inferencja modeli** — Zbudowałem pipeline fine-tuningu LoRA/DoRA dla modeli postaci (wersjonowane snapshoty datasetów, captioning przez LLM, rejestr checkpointów). Uruchamiałem modele wideo open-weight do 33 mld parametrów na kartach 32 GB dzięki kwantyzacji int8/NVFP4 i limitom pamięci dobieranym do obciążenia. Modele i parametry wybierałem na podstawie benchmarków i testów A/B, a nie na wyczucie.
- Mierzalna jakość** — Zaprojektowałem metryki i zestawy testów, które wyłapują przypadki brzegowe: dryf tożsamości mierzony embeddingami twarzy, dokładność punktów pozy, zgodność obrazu co do piksela. Kontrolowane testy A/B obniżyły odsetek klątek z krytyczną wadą tożsamości z 89% do 0%, a dopracowanie briefu podniosło odsetek udanych klipów na nowym silniku z 36% do 69%.
- Wydajność i koszty** — Przeniosłem generowanie wideo z płatnych API na modele uruchamiane we własnej puli GPU w chmurze (szczytowo 20–26 kart): koszt obliczeń na film niższy ~6x, wydatki na API niższe o 89% w ciągu miesiąca. Skróciłem preprocessing pozy z 12 minut na CPU do 45 sekund na GPU, po namierzeniu konfliktu pakietów ONNX Runtime CPU/GPU. Przyspieszenie renderingu o 33–36% trafiło na produkcję dopiero wtedy, gdy test regresji na GPU potwierdził identyczny obraz co do piksela (SSIM 1,0).
- Rozproszone przetwarzanie zadań** — Zbudowałem trwałą kolejkę zadań (durable job queue: leases, heartbeats, fencing tokens, retry, klasy zasobów GPU) oraz dispatcher oparty na centralnym rejestrze, który zastąpił zawodne przetwarzanie wsadowe. Efekt: 230 klipów wyrenderowanych w jedną noc na maks. 12 kartach RTX 5090 oraz seria 380 klipów na maks. 20 GPU bez ani jednego błędu kolejki, przy przestojach poniżej 1%.
- Stabilna produkcja i bezpieczne wdrożenia** — 951 filmów w ok. 30 godzin bez ani jednego nieudanego renderu. Każdy release przechodzi przez CI z 1145 testami automatycznymi; rollback jedną komendą i blokada deploymentu w trakcie renderingu.
- Zespół i biznes** — Jednostkowy koszt zmienny spadł o 45% m/m przy wzroście produkcji o 18%. Rekrutowałem i prowadziłem 2–4 operatorów (wydajność na roboczogodzinę x3,6 w ciągu miesiąca); przygotowałem cennik, pakiety usług i prezentację sprzedażową — po czym spłynęły pierwsze zamówienia od klientów zewnętrznych.

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

Head of Content Production Kemp Center / kursyszkolenia.online (e-learning) 12.2024 – 03.2026 · Poznań

- ◆ **4,5x szybsza produkcja, ok. 600 wykładów** — Zamiast liniowego procesu (jeden montażysta — jeden wykład) wprowadziłem produkcję równoległą: czas finalizacji wykładu spadł z 3 godzin do 40 minut. Od zera zbudowałem 4-osobowy zespół, który oddawał 3–4 wykłady dziennie w 20 kursach.
- ◆ **AI i automatyzacja w produkcji** — Tła generowane w Midjourney, animowany B-roll (Runway, Kling), wirtualni prezenterzy z syntetycznym głosem (HeyGen, ElevenLabs) i automatyzacja wsadowa w ExtendScript; zautomatyzowana migracja 221 archiwalnych wykładów bez utraty danych.

Claude / GPT API Midjourney Runway Kling HeyGen ElevenLabs ExtendScript

Freelance Video Producer & Motion Designer Fiverr · Top Rated 10.2022 – 12.2024 · zdalnie

- ◆ **298 projektów, 120 klientów z USA** — Przekształciłem jednoosobową działalność w mikroagencję z podwykonawcami: od briefu i koncepcji, przez zdjęcia, montaż, color grading i VFX, po materiały na różne platformy; 70% realizacji bez poprawek.

Reżyser / Lead Video Producer Państwowy Zespół Ludowy Pieśni i Tańca „Mazowsze” 2021 – 2022 · Otrębusy

- ◆ Kompleksowa produkcja wideo dla narodowego zespołu (170+ artystów): ponad 200 koncertów w Polsce i za granicą, realizacje wielokamerowe na żywo, spoty TV i animowane reklamy na Dworcu Centralnym i w komunikacji miejskiej, obsługa wydarzeń z udziałem prezydenta i korpusu dyplomatycznego.

Video Producer i operator kamery Rozmarin Studio / Automax Detailing 2015 – 2021 · Warszawa i Kijów

- ◆ Sześć lat produkcji dla branży motoryzacyjnej, reklamowej i kreatywnej — od operatora kamery do szefa studia.

WYBRANE PROJEKTY

Max Chronicle open source · MIT

Lokalna pamięć dla agentów AI: serwer MCP, CLI i baza zdarzeń w SQLite. Wyszukiwanie hybrydowe (pełnotekstowe, wektorowe na lokalnych embeddingach, temporalne) z własnym zestawem ewaluacji trafności; ponad 700 testów w CI.

Python MCP Retrieval Ewaluacja

LocalFlow open source · MIT

Lokalny pipeline mowy i LLM dla macOS: dyktowanie i transkrypcja spotkań wyłącznie na modelach lokalnych, rozpoznawanie mówców i przeszukiwalne archiwum z odnośnikami do źródeł; skrypty do ewaluacji modeli. 97 testów, CI na macOS.

Swift MLX Lokalne LLM Mowa

RZMRN Digest autonomiczny pipeline LLM

15-etapowy pipeline LLM zasilany z 84 źródeł: zbieranie, deduplikacja, równoległa analiza, wykrywanie anomalii, tłumaczenie i publikacja dwa razy dziennie, z watchdogiem i alertami. Przez miesiące działał bez nadzoru.

Python Pipeline'y LLM Cloudflare

UMIEJĘTNOŚCI

Agenci AI i automatyzacja	Claude Code, Codex, systemy wieloagentowe z limitami budżetu i guardrailami, serwery MCP, tool calling, structured outputs, prompt engineering, RAG, pamięć agentów (mem0, Kuzu, embeddingi)
Modele i trening	Fine-tuning LoRA/DoRA (ai-toolkit), inferencja modeli open-weight, kwantyzacja (int8, NVFP4, fp8), optymalizacja pamięci i kerneli attention, modele dyfuzyjne obrazu i wideo (WAN, LTX, MiniMax H3, FLUX, Z-Image), computer vision (ViTPose, RTMPose, SAM, embeddingi twarzy), LLM/VLM (Claude, GPT, DeepSeek, Qwen-VL), mowa (Whisper, pyannote)
Ewaluacja i jakość	Projektowanie metryk, zestawy testów i przypadki brzegowe, testy A/B z jedną zmienną, testy regresji i zgodności, zestawy ewaluacyjne, human-in-the-loop
Infrastruktura i narzędzia	Python, FastAPI, SQLite (WAL), trwałe kolejki zadań, Docker (obrazy CUDA), CI/CD w GitHub Actions, pula GPU na RunPod, ComfyUI, Linux, Next.js
Zarządzanie i realizacja	Prowadzenie produkcyjnej platformy AI, specyfikacje i kryteria akceptacji, współpraca z klientem, unit economics, budowa i prowadzenie zespołów
Produkcja wideo (10+ lat)	Premiere Pro, After Effects, DaVinci Resolve, Cinema 4D, color grading, motion design, realizacja wielokamerowa na żywo, dron (EASA A1/A2/A3)

WYKSZTAŁCENIE I CERTYFIKATY

Zarządzanie informacją Polsko-Japońska Akademia Technik Komputerowych (PJATK) 2016 – 2019

Przedmioty: architektura baz danych, projektowanie API, inżynieria procesów. Certyfikaty: Cinema 4D, dron EASA, color grading.

Wyrażam zgodę na przetwarzanie moich danych osobowych w celu rekrutacji zgodnie z RODO (Rozporządzenie UE 2016/679).