



Ihor Kostyniuk



Senior AI Automation Engineer | LLM Agents & Orchestration, MCP, RAG, Voice AI

Senior AI Automation Engineer, 3+ years shipping production LLM systems on self-hosted infrastructure. I run a 7-channel conversational AI support platform, AI video pipelines for 22 YouTube creators and 20+ client channels (45-65 minute films at \$22-33 each in compute), and 10+ production applications of my own. I write the orchestration layer myself instead of adopting a framework: tool-calling loops, custom MCP servers, multi-provider routing with automatic fallback, checkpointed pipelines that resume after a crash, human-in-the-loop gates, and an LLM-as-judge eval and regression harness. I ship end to end: multi-tenant SaaS, native Android and desktop applications, AI video pipelines, and recruiting systems built to EU AI Act and GDPR obligations.

EMAIL

koss.flow@gmail.com

PHONE

[+380679368855](tel:+380679368855)

LOCATION

Odesa, Ukraine

LINK

<https://linkedin.com/in/koss-flow>

LINK

https://t.me/koss_flow

LINK

<https://demo.kossflow.com>

01 Experience

Generative AI Engineer

FEB 2024 → NOV 2024

YT-Wizards

- Scaled AI video production to 22 YouTube creators and 20+ client channels by forking reusable pipeline skeletons per niche (war and aviation history, space, finance, watches, lore, travel documentaries); many are handed over and run self-serve on the creator's own machine or server.
- Built the orchestration behind renders that run for hours per job: checkpointed state so any stage resumes after a crash, per-step retry with backoff, and circuit breakers around each model provider.
- Built the FFmpeg filter_complex engine that renders those 45-60 minute films (depth-map parallax, word-timed captions), and the handover layer that moved rendering to the clients: bundled installer, deploy-key provisioning, one-click update, server-side render and frame-level editing in the UI. Non-technical clients now run their own renders on their own hardware instead of every video going through me.
- Delivered 45-65 minute documentaries (~235 scenes in a 45-60 minute film) at \$22-33 per video in compute, with unit economics measured from a \$0.16/min budget mode up to premium tiers.
- Gated output quality with models rather than review by eye: Gemini vision identity verification on photos, text-artifact detection with automatic regeneration on AI images, and freezedetect frame QA.

Senior AI Automation Engineer

SEP 2024 → OCT 2024

SOKIL

- Own a 7-channel conversational AI support platform (WhatsApp, Telegram, Viber, Instagram, Facebook, Web, Email) end to end: agent design, tool calling, per-tenant Qdrant RAG, evals and deployment. Cleared Meta App Review for production messaging permissions.
- Turned the embeddable web widget into a lead channel for the same 7-channel agent: it runs on customer sites in the host page's language and theme, is screen-reader accessible, and writes a visitor who leaves a phone number straight into the CRM as a card plus an outbound webhook, so leads reach downstream systems without anyone re-typing them.
- Shipped a multi-tenant AI recruiting CRM (33 database migrations, RLS on every table, subdomain-per-workspace) running ~20 AI tasks including CV parsing, few-shot-calibrated ATS scoring, outreach generation, and vacancy parsing, fed by a 17-source job-scraping pipeline with HMAC callbacks and IMAP ingestion.
- Delivered EU AI Act compliance for high-risk recruiting AI: DPIA/DPA templates, self-serve DSAR export and erasure, human-in-the-loop score overrides with audit trail, and a bias-monitoring dashboard (4/5 rule cohort analysis), backed by 86+ Playwright security tests and CI/CD with smoke-tested deploys.
- Made AI provider choice a per-tenant setting: independent chat and embedding providers across three vendors, a live model catalogue with current prices, and graceful handling when a provider drops a model an organization still has selected.

AI Automation Engineer

AUG 2025 → APR 2026

PICHE, LTD

- Audited four business functions (procurement, accounting, legal, recruitment), turned the bottlenecks into implementation specs, and then built the systems those specs described.
- Built a custom ATS and recruitment pipeline: vacancies and candidate data from major and niche job boards land in SmartSuite as one system of record, then CVs are parsed, candidates scored and the technical interview generated, so screening stopped being a manual step for the recruiters.
- Architected a LegalTech document-generation system with docxtemplater and Gonenberg: contracts and management reports are assembled from reusable templates, so drafting moved from copy-editing old documents to filling a form.
- Built the webhook-driven operations app the teams work from: task scheduling, run monitoring, work-hour tracking and operating-expense control in one view, so schedules and spend came out of the system instead of being assembled by hand for each report.
- Ran the infrastructure under all of it: Docker on VPS with n8n, OpenClaw agents and shared internal libraries, with uptime monitoring and error tracking wired into every automation, so a broken flow raised an alert instead of being found by the team that depended on it.

AI Automation Engineer

SEP 2023 → PRESENT

KossFlow

- Operate 10+ production applications on a self-hosted VPS: Next.js, Supabase, Docker Compose, Caddy, centralised monitoring and nightly backups.
- Built KossOps, a single view over 9 servers with 36 watchdogs, a dead-man switch that reports when the monitor itself dies, and a native Android client with push alerts.
- Built a self-hosted video-meeting platform on LiveKit with an in-call voice AI assistant (wake-word gated, RAG over the organisation knowledge base), speaker-attributed transcription with an STT fallback chain (Groq Whisper, Deepgram, self-hosted), and automatic meeting-to-specification PDF artefacts.
- Practice adversarial security engineering on my own systems: multi-agent audits found and closed a remote-code-execution chain (a public anon key reaching a job queue executed as root) and a stored-XSS bypass using entity-obfuscated payloads and tag splitting - neither was reported by gitleaks, semgrep or trivy, because the risk lived in the combination rather than in any single file.
- Built an on-device AI stack on Intel Arc: 21 video courses (1,300+ lessons, 170+ hours) transcribed with OpenVINO Whisper and indexed into Qdrant (19k chunks, bge-m3) behind a custom MCP server that gives coding agents instant course knowledge, plus a voice-cloning studio with zero-shot Russian, Ukrainian and English cloning, best-of-N generation gated by ASR accuracy and speaker-similarity scoring, and an OpenAI-compatible TTS API.
- Run an autonomous multi-channel content system (LinkedIn, Instagram, Telegram, Threads) with a Telegram approval bot, 18-source deep fact-checking, cross-model verdict verification (GPT vs Claude), and multi-provider LLM routing with automatic fallback that keeps daily cost near zero.
- Shipped as products rather than demos: a native Android personal-finance app (Kotlin, Jetpack Compose, agentic AI chat over the user's own finances, entrepreneur tax calendar, Wear OS companion) released and in daily use at v5.13; Papir Desktop (Tauri 2, React 19, TypeScript strict), a fully offline PDF suite covered by 3,164 unit tests and 169 Rust tests that replaced a commercial reader in daily use; and a self-hosted over-the-air update channel with signed artefacts that took the cable out of every Android release.

03 | Skills

LLM Engineering AI Agents Agentic AI Agent Orchestration Multi-Agent Systems Multi-Step Reasoning

Tool Calling & Function Calling MCP (Model Context Protocol) Server Development

Agent Memory & Context Management Prompt Engineering & Optimization LLM Evaluation (evals)

Custom Eval Harness LLM-as-Judge Verification Regression Testing Human-in-the-Loop Systems

Retry / Circuit-Breaker / Fallback Design RAG (Retrieval-Augmented Generation) Agentic RAG

Vector Databases Qdrant pgvector Embeddings Semantic Search Conversational AI Chatbots

Voice Agents (LiveKit) Speech-to-Text (Whisper) Anthropic Claude OpenAI Google Gemini Groq

Multi-Provider Routing Python TypeScript Node.js FastAPI Next.js React PostgreSQL SQL

Supabase REST APIs Docker Linux CI/CD n8n (self-hosted, queue mode) Multi-Tenant Isolation

Row-Level Security (RLS) EU AI Act Technical Controls GDPR (DSAR / Bias Monitoring)

Spec-Driven Agent Design Context Engineering Prompt / Config Versioning

Automated Fact-Checking Pipelines Observability (Sentry / GlitchTip) OpenVINO / On-Device Inference

Text-to-Speech Zero-Shot Voice Cloning Speaker-Attributed Transcription FFmpeg filter_complex

Kotlin / Jetpack Compose Flutter Tauri 2 Ansible Caddy Playwright E2E OAuth 2.0 / JWT

AES-256-GCM Encryption (BYOK) HMAC Webhook Verification Adversarial Security Review WhatsApp Cloud API

Meta Login for Business OTA App Distribution Redis WebSockets / SSE

04 | Languages

Claude Code Mastery: Basics, Web Projects & Agents Udemy Arnold Oberleiter	JUL 2026
Vibe Coding — Build Apps with Claude Code, Cursor, and Codex Udemy Arnold Oberleiter	MAY 2026
OpenClaw: Run Powerful & Autonomous AI Agents Securely Udemy Arnold Oberleiter	APR 2026
MCP: Building Agents with Claude, Cursor, Flowise, Python & n8n Udemy Arnold Oberleiter	FEB 2026
AI Voice Agents: Automation with Vapi, ElevenLabs, n8n & MCP Udemy Arnold Oberleiter	DEC 2025
AI Automation: Build LLM Apps & AI Agents with n8n & APIs Udemy Arnold Oberleiter	SEP 2025
Automation #withMake Web3nity	JAN 2025