



Ігор Костинюк

Senior AI Automation Engineer | LLM-агенти та оркестрація, MCP, RAG, Voice AI



Senior AI Automation Engineer, 3+ роки досвіду запуску продакшн-систем LLM на власній інфраструктурі. Веду 7-канальну розмовну платформу підтримки з ШІ, ШІ-відеопайплайни для 22 YouTube-авторів і 20+ клієнтських каналів (фільми по 45-65 хвилин за \$22-33 обчислень кожен), і 10+ власних продакшн-застосунків. Сам пишу шар оркестрації замість упровадження фреймворку: цикли tool calling, власні сервери MCP, маршрутизацію між кількома провайдерами з автоматичним фолбеком, пайплайни з контрольними точками, що відновлюються після збою, гейти з участю людини, а також систему evals і регресійних перевірок за схемою LLM-as-judge. Запускаю продукти від початку до кінця: мультитенантний SaaS, нативні Android- і настільні застосунки, відеопайплайни з ШІ та рекрутингові системи, побудовані з урахуванням вимог EU AI Act і GDPR.

EMAIL

koss.flow@gmail.com

PHONE

[+380679368855](tel:+380679368855)

LOCATION

Одеса, Україна

LINK

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

LINK

https://t.me/koss_flow

LINK

<https://demo.kossflow.com>

01 Experience

Generative AI Engineer

FEB 2026 → NOV 2026

YT-Wizards

- Масштабував виробництво відео з ШІ до 22 YouTube-авторів і 20+ клієнтських каналів на багаторазових каркасах пайплайнів, адаптованих під кожну нішу; багато з них передані авторам і працюють у них самостійно.
- Побудував оркестрацію для рендерів, що виконуються годинами на одне завдання: стан з контрольними точками, щоб будь-який етап відновлювався після збою, ретраї кожного кроку з backoff і circuit breakers навколо кожного модельного провайдера.
- Побудував рушій FFmpeg filter_complex, який рендерить ці фільми тривалістю 45-60 хвилин (паралакс за картою глибини, субтитри з таймінгом за словами), і шар передавання роботи клієнтам, який переніс рендеринг до клієнтів: інстальатор у комплекті, видача deploy-key, оновлення одним кліком, серверний рендеринг і покадрове редагування в інтерфейсі. Нетехнічні клієнти тепер запускають власні рендери на власному обладнанні, замість того щоб кожне відео проходило через мене.
- Поставив документальні фільми тривалістю 45-65 хвилин (близько 235 сцен у фільмі на 45-60 хвилин) за \$22-33 обчислювальних витрат на відео, з юніт-економією від бюджетного режиму \$0.16/хв до преміальних рівнів.
- Контролював якість результату моделями, а не переглядом вручну: перевірка ідентичності на фото через Gemini vision, виявлення текстових артефактів з автоматичною регенерацією зображень ШІ та покадровий контроль якості через freezedetect.

Senior AI Automation Engineer

SEP 2024 → OCT 2026

SOKIL

- Відповідаю за 7-канальну розмовну платформу підтримки з ШІ (WhatsApp, Telegram, Viber, Instagram, Facebook, веб, електронна пошта) від початку до кінця: дизайн агентів, tool calling, Qdrant RAG на рівні тенанта, evals і розгортання. Пройшов Meta App Review для продакшн-дозволів на обмін повідомленнями.
- Перетворив вбудований вебвіджет на канал залучення лідів для того самого 7-канального агента: він працює на сайтах клієнтів мовою й темою сторінки, доступний для зчитувачів екрана й записує відвідувача, який залишив номер телефону, прямо в CRM як картку плюс вихідний вебхук, щоб ліди потрапляли в подальші системи без ручного передруку.
- Запустив мультитенантну CRM для рекрутингу з ШІ (33 міграції бази даних, RLS на кожній таблиці, піддомен на кожен робочий простір), що виконує приблизно 20 завдань ШІ, зокрема парсинг резюме, оцінювання ATS, відкаліброване few-shot, генерацію ауричу та парсинг вакансій, із даними з 17-джерельного пайплайну скрапінгу вакансій із HMAC-зворотними викликами та отриманням через IMAP.
- Забезпечив відповідність EU AI Act для високоризикового ШІ в рекрутингу: шаблони DPIA/DPA, самостійний експорт і видалення DSAR, перевизначення оцінок з участю людини й аудиторським слідом, а також дашборд моніторингу упередженості (когортний аналіз за правилом 4/5), підкріплені 86+ безпековими тестами Playwright і CI/CD з розгортаннями, перевіреними smoke tests.
- Зробив вибір провайдера ШІ налаштуванням на рівні тенанта: незалежні провайдери чату й embeddings у трьох постачальниках, живий каталог моделей з актуальними цінами та коректна обробка ситуацій, коли провайдер вилучає модель, яку організація все ще має вибраною.

PICHE, LTD

- Проаудював чотири бізнес-функції (закупівлі, бухгалтерію, юридичний напрям, рекрутинг), перетворив вузькі місця на специфікації впровадження, а потім побудував системи, описані цими специфікаціями.
- Побудував власну ATS і рекрутинговий пайплайн: вакансії та дані кандидатів із великих і нішевих сайтів працевлаштування потрапляють у SmartSuite як єдине джерело правди, після чого резюме парсяться, кандидати оцінюються й генерується технічна співбесіда, тож скринінг перестав бути ручним кроком для рекрутерів.
- Спроектував систему генерації юридичних документів із docxtemplater і Gonenberg: договори та управлінські звіти збираються з багаторазових шаблонів, тож підготовка документів перейшла від редагування старих файлів до заповнення форми.
- Побудував операційний застосунок на основі вебхуків, у якому працюють команди: планування завдань, моніторинг запусків, облік робочого часу й контроль операційних витрат в одному поданні, тож графіки та витрати виходили із системи, а не збиралися вручну для кожного звіту.
- Забезпечував інфраструктуру під усім цим: Docker на VPS з n8n, агентами OpenClaw і спільними внутрішніми бібліотеками, з моніторингом доступності та відстеженням помилок, підключеними до кожної автоматизації, щоб зламаний потік створював сповіщення, а не виявлявся командою, яка від нього залежала.

AI Automation Engineer

KossFlow

- Експлуатую 10+ продакшн-застосунків на власному VPS: Next.js, Supabase, Docker Compose, Caddy, централізований моніторинг і нічні резервні копії.
- Побудував KossOps, єдине подання для 9 серверів із 36 сторожами, dead-man switch, який повідомляє, коли сам монітор перестає працювати, і нативний Android-клієнт із push-сповіщеннями.
- Побудував власну платформу відеозустрічей на LiveKit із голосовим ШІ-асистентом у дзвінку (активація ключовим словом, RAG над базою знань організації), транскрипцією з прив'язкою до мовців із ланцюжком STT-фолбеків (Groq Whisper, Deepgram, власний хостинг) та автоматичними PDF-артефактами, що перетворюють зустрічі на специфікації.
- Практикую змагальну безпекову інженерію на власних системах: багатоагентні аудити знайшли й закрили ланцюжок віддаленого виконання коду (публічний анонімний ключ діставався черги завдань, що виконувалася як root) і обхід збереженого XSS через навантаження з обфускацією сутностей і розбиттям тегів - жоден із них не був виявлений gitLeaks, semgrep або trivy, бо ризик жив у поєднанні, а не в окремому файлі.
- Побудував локальний стек ШІ на Intel Arc: 21 відеокурс (1300+ уроків, 170+ годин) транскрибовано через OpenVINO Whisper і проіндексовано в Qdrant (19 тис. фрагментів, bge-m3) за власним MCP-сервером, який дає агентам кодування миттєві знання з курсів, плюс студія клонування голосу з zero-shot-клонуванням російською, українською та англійською, генерацією best-of-N через гейт за точністю ASR і оцінкою схожості мовця, а також OpenAI-сумісним TTS API.
- Веду автономну мультиканальну систему контенту (LinkedIn, Instagram, Telegram, Threads) з Telegram-ботом затвердження, глибокою перевіркою фактів за 18 джерелами, перевіркою вердиктів між моделями (GPT проти Claude) і маршрутизацією LLM між кількома провайдерами з автоматичним фолбеком, що тримає щоденні витрати близькими до нуля.
- Запустив як продукти, а не демонстрації: нативний Android-застосунок для особистих фінансів (Kotlin, Jetpack Compose, агентний чат ШІ над власними фінансами користувача, податковий календар підприємця, супутній застосунок Wear OS), випущений і щоденно використовуваний у версії v5.13; Papir Desktop (Tauri 2, React 19, TypeScript у строгому режимі), повністю офлайн-набір для роботи з PDF, покритий 3164 модульними тестами та 169 тестами Rust, який замінив комерційний PDF-рідер у щоденному використанні; а також власний канал бездротових оновлень із підписаними артефактами, який прибрав кабель із кожного Android-релізу.

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

04

Languages

Англійська	C1	Українська	C2	Російська	РІДНА
------------	----	------------	----	-----------	-------

06

Courses

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