



Игорь Костынюк

Senior AI Automation Engineer | LLM-агенты и оркестрация, MCP, RAG, Voice AI



Senior AI Automation Engineer, 3+ года опыта в выпуске продакшн LLM-систем на собственной инфраструктуре. Веду 7-канальную разговорную ИИ-платформу поддержки, ИИ-видеопаплайны для 22 YouTube-авторов и 20+ клиентских каналов (фильмы по 45-65 минут за \$22-33 вычислений каждый), и 10+ собственных продакшн-приложений. Сам пишу слой оркестрации вместо внедрения фреймворка: циклы tool calling, кастомные MCP-серверы, маршрутизацию между несколькими провайдерами с автоматическим фолбэком, пайплайны с контрольными точками, возобновляемые после сбоя, гейты с участием человека, а также стенд LLM-as-judge для оценок и регрессий. Выпускаю продукты от начала до конца: мультитенантный SaaS, нативные Android- и настольные приложения, ИИ-видеопаплайны и рекрутинговые системы, построенные с учетом требований EU AI Act и GDPR.

EMAIL

koss.flow@gmail.com

PHONE

+380679368855

LOCATION

Одесса, Украина

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

- Масштабировал AI-видеопроизводство до 22 YouTube-авторов и 20+ клиентских каналов на переиспользуемых каркасах пайплайнов, адаптированных под каждую нишу; многие переданы авторам и работают у них сами.
- Построил оркестрацию для рендеров, которые выполняются часами на одну задачу: состояние с контрольными точками, чтобы любой этап возобновлялся после сбоя, ретраи на каждом шаге с экспоненциальной задержкой и circuit breakers вокруг каждого провайдера моделей.
- Построил движок FFmpeg filter_complex, который рендерит эти фильмы длительностью 45-60 минут: параллакс по карте глубины, субтитры с привязкой к словам, а также слой передачи рендеринга клиентам: установщик в комплекте, выдача deploy-ключей, обновление в один клик, серверный рендер и пок кадровое редактирование в интерфейсе. Нетехнические клиенты теперь запускают собственные рендеры на своем оборудовании, вместо того чтобы каждое видео проходило через меня.
- Выпустил документальные фильмы длительностью 45-65 минут (около 235 сцен в фильме на 45-60 минут) со стоимостью вычислений \$22-33 за видео и экономикой от бюджетного режима \$0.16/мин до премиальных уровней.
- Ограничил качество вывода моделями, а не визуальной проверкой: проверка идентичности на фото через Gemini vision, обнаружение текстовых артефактов с автоматической регенерацией AI-изображений и пок кадровая проверка качества через freeze detect.

Senior AI Automation Engineer

SEP 2024 → OCT 2026

SOKIL

- Отвечаю за 7-канальную разговорную AI-платформу поддержки (WhatsApp, Telegram, Viber, Instagram, Facebook, веб, электронная почта) от начала до конца: дизайн агентов, tool calling, Qdrant RAG на уровне тенанта, evals и развертывание. Прошел Meta App Review для продакшн-разрешений на обмен сообщениями.
- Превратил встраиваемый веб-виджет в канал лидогенерации для того же 7-канальной агента: он работает на сайтах клиентов на языке и в теме страницы-хоста, доступен для программ чтения с экрана и записывает посетителя, оставившего номер телефона, прямо в CRM как карточку плюс исходящий вебхук, чтобы лиды доходили до нижестоящих систем без ручного перепечатывания.
- Выпустил мультитенантную AI CRM для рекрутинга: 33 миграции базы данных, RLS на каждой таблице, отдельный поддомен на рабочее пространство; система выполняет около 20 AI-задач, включая разбор CV, калиброванную few-shot оценку ATS, генерацию аурича и разбор вакансий, и питается пайплайном сбора вакансий из 17 источников с HMAC-колбэками и приемом через IMAP.
- Обеспечил соответствие EU AI Act для высокорискового рекрутингового AI: шаблоны DPIA/DPA, самостоятельный экспорт и удаление по DSAR, переопределение оценок с участием человека и аудиторским следом, а также панель мониторинга смещений: когортный анализ по правилу 4/5, подкрепленный 86+ тестами безопасности Playwright и CI/CD с развертываниями, проверенными smoke tests.
- Сделал выбор AI-провайдера настройкой на уровне тенанта: независимые провайдеры чата и embeddings у трех поставщиков, живой каталог моделей с актуальными ценами и мягкая обработка ситуации, когда провайдер удаляет модель, все еще выбранную организацией.

PICHE, LTD

- Провел аудит четырех бизнес-функций: закупок, бухгалтерии, юридического блока и рекрутинга; превратил узкие места в спецификации реализации, а затем построил системы, описанные в этих спецификациях.
- Построил кастомную ATS и рекрутинговый пайплайн: вакансии и данные кандидатов с крупных и нишевых сайтов вакансий попадают в SmartSuite как единую систему учета, затем CV разбираются, кандидаты оцениваются и генерируется техническое собеседование, так что скрининг перестал быть ручным этапом для рекрутеров.
- Спроектировал LegalTech-систему генерации документов на docxtemplater и Gotenberg: договоры и управленческие отчеты собираются из переиспользуемых шаблонов, поэтому подготовка документов перешла от копирования и правки старых файлов к заполнению формы.
- Построил операционное приложение на вебхуках, из которого работают команды: планирование задач, мониторинг запусков, учет рабочего времени и контроль операционных расходов в одном представлении, так что графики и затраты стали формироваться системой, а не собираться вручную для каждого отчета.
- Вел инфраструктуру под всем этим: Docker на VPS с n8n, агентами OpenClaw и общими внутренними библиотеками; мониторинг доступности и отслеживание ошибок подключены к каждой автоматизации, поэтому сломанный поток поднимал оповещение, а не обнаруживался зависимой от него командой.

AI Automation Engineer

KossFlow

- Эксплуатирую 10+ продакшн-приложений на собственном VPS: Next.js, Supabase, Docker Compose, Caddy, централизованный мониторинг и ночные резервные копии.
- Построил KossOps: единое представление по 9 серверам с 36 сторожами, dead-man switch, который сообщает, когда умирает сам монитор, и нативный Android-клиент с push-оповещениями.
- Построил собственную платформу видеовстреч на LiveKit с голосовым AI-ассистентом внутри звонка: активация по ключевому слову, RAG по базе знаний организации; транскрибация с атрибуцией говорящего и цепочкой STT-фолбэков: Groq Whisper, Deepgram, собственная развернутая модель; а также автоматические PDF-артефакты из встречи в спецификацию.
- Практикую состоятельную инженерию безопасности на собственных системах: мультиагентные аудиты нашли и закрыли цепочку удаленного выполнения кода: публичный анонимный ключ получал доступ к очереди задач, выполняемой с правами суперпользователя, а также обход сохраненного межсайтового скриптинга через нагрузки с замаскированными сущностями и разбиением тегов; ни gitleaks, ни semgrep, ни trivy этого не сообщили, потому что риск жил в комбинации, а не в отдельном файле.
- Построил локальный ИИ-стек на Intel Arc: 21 видеокурс, 1300+ уроков (170+ часов), транскрибированы через OpenVINO Whisper и проиндексированы в Qdrant: 19 тыс. фрагментов, bge-m3; поверх этого работает кастомный MCP-сервер, дающий агентам для кодига мгновенный доступ к знаниям курсов, плюс студия клонирования голоса с zero-shot клонированием на русском, украинском и английском, best-of-N генерацией, ограниченной точностью ASR и оценкой сходства говорящего, и OpenAI-совместимым TTS API.
- Веду автономную мультиканальную контент-систему: LinkedIn, Instagram, Telegram, Threads; с Telegram-ботом для согласования, глубокой проверкой фактов по 18 источникам, кросс-модельной проверкой вердиктов: GPT против Claude; и маршрутизацией LLM через нескольких провайдеров с автоматическим фолбэком, удерживающей ежедневные расходы около нуля.
- Выпустил именно продукты, а не демо: нативное Android-приложение для личных финансов: Kotlin, Jetpack Compose, агентный AI-чат по собственным финансам пользователя, налоговый календарь предпринимателя, компаньон Wear OS; выпущено и ежедневно используется на версии v5.13; Papir Desktop: Tauri 2, React 19, строгий TypeScript, полностью офлайн-набор для работы с PDF, покрытый 3,164 модульными тестами и 169 Rust-тестами, который заменил коммерческую программу для чтения в ежедневном использовании; а также собственный канал обновлений по воздуху с подписанными артефактами, убравший кабель из каждого Android-релиза.

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

04

Languages

Английский	C1	Украинский	C2	Русский	Родной
------------	----	------------	----	---------	--------

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