Oleksandr Pokariev

AI Engineer | Applied LLM Systems · RAG · Voice & Document AI · LLM Ops

📍 Ostrówek, near Warsaw, Poland (05-205) — remote, timezone GMT+1/+2 (fits GMT-2…GMT+5)
Email: pokarev@gmail.com
Phone: +48 534 998 798
GitHub: https://github.com/grabbly
LinkedIn: https://www.linkedin.com/in/alex-pokariev-586959356
English: B1 · Availability: open to immediate start

Professional Summary

Software engineer with 15+ years of commercial development, the last 2+ years focused full-time on **applied AI / LLM systems**. I design and operate production AI services around Large Language Models: RAG pipelines, real-time voice agents, document/OCR extraction, and multi-provider LLM integration with routing and fallback.

I work the way this role is framed — **mixing ready-made tools with custom-built solutions**: orchestrating pipelines in n8n/LangChain nodes where that ships fastest, and writing native code (Python/Node) where control and reliability matter. Strong on end-to-end ownership: technical design → implementation → containerized deployment → post-deployment monitoring and auto-healing. Primary language **Python (FastAPI)**, plus Node/TypeScript, PHP and C#.

Core Competencies (mapped to the role)

- **LLM pipelines & wrappers:** model selection and configuration, provider-routing with automatic fallback, self-healing model watchdogs, OpenAI-compatible façades over self-hosted backends
- **RAG:** embeddings, chunking, top-k retrieval, per-tenant namespaces; production RAG on **Qdrant** (orchestrated via n8n/LangChain nodes) **and** a hand-built vector-retrieval memory store (embeddings + cosine + context injection) written from scratch in Node — no framework
- **OCR / document AI:** vision-LLM OCR and **structured entity extraction** from business documents (receipts/invoices) — key-field extraction, JSON schema output, validation and multi-sample verification
- **Voice AI:** real-time speech pipelines — Whisper STT, Silero/Piper/ElevenLabs TTS, WebSocket streaming, Twilio telephony
- **Prompt engineering:** dedicated prompt-construction modules (persona/facts/summary/rules), structured-output prompting with validation & retry, latency fast-lane variants
- **Evaluation / experimentation:** custom benchmark bench comparing prompts and model configs by latency, token cost and output quality, with per-call model-override for A/B comparison
- **LLM Ops / infrastructure:** self-hosting and serving open-source models (**vLLM**, Ollama — Qwen2.5-7B-AWQ, Llama-family, gemma), Docker, Linux, Nginx, PM2, health-monitoring + auto-heal
- **Delivery:** Git, CI/CD via webhook auto-deploy, Docker, iterative delivery, production ownership

Technology Stack

- **Languages:** Python, JavaScript/TypeScript, C#, PHP
- **Backend / APIs:** FastAPI, Node.js (Fastify/Express), REST, WebSockets (real-time), SSE streaming
- **LLM providers:** OpenAI (Chat/Assistants/Realtime/Whisper), Anthropic (Claude API), Google Gemini (vision), DeepSeek, Mistral; self-hosted Qwen/Llama/gemma via vLLM & Ollama

- **RAG / vector:** Qdrant (production, n8n-orchestrated), custom embeddings + cosine vector store; LangChain (via n8n orchestration nodes)
- **Voice / vision:** Whisper, faster-whisper, Silero, Piper, ElevenLabs; Gemini Vision (OCR / receipt & document extraction)
- **Orchestration / automation:** n8n, LangFlow, LangGraph, custom workflows
- **Databases:** MySQL/MariaDB, PostgreSQL, Prisma; SQL data validation/transforms
- **DevOps / Cloud:** Docker, Linux, Nginx, PM2, WireGuard; AWS, GCP, Azure
- **Tools:** Git, Postman, Jira, pgAdmin/phpMyAdmin, Nginx Proxy Manager

Professional Experience

Founder & Lead AI Developer — AceBox (privacy-first AI systems)

2024 – Present · https://acebox.eu

- Designed and shipped **production LLM services** for business automation across 5+ web projects
- Built and operate a **RAG pipeline on Qdrant** (embeddings, chunking, per-tenant namespaces) orchestrated via n8n/LangChain nodes; hand-built a separate **native vector-retrieval memory store** (embeddings + cosine + top-k + context injection) in Node
- Delivered **real-time AI voice agents** (Whisper STT + Silero/Piper/ElevenLabs TTS + WebSocket + Twilio), including a multi-tenant voice-agent SaaS backend
- Built a **document/OCR extraction service** (vision LLM): verbatim OCR + structured JSON extraction from receipts/invoices with field validation
- Implemented **multi-provider LLM integration** with provider-routing, automatic fallback and a self-healing model watchdog that fails over when a model is retired
- Self-hosted and operated open-source models with **vLLM** and Ollama (Qwen2.5-7B-AWQ, Llama-family, gemma) on private Linux infrastructure
- Own the full lifecycle: technical design → Docker/Linux deployment → **monitoring & auto-heal** (health checks, container auto-restart, PM2)

Senior / Lead Developer (C#, Unity) — production products

2017 – 2024

- 10+ years of large-scale production software with focus on performance optimization, reliability and rapid prototyping; shipped 100+ mobile projects. Foundation for current AI-systems engineering.

Selected AI Projects

- **Twinlix — AI Voice Agent Platform** (formerly DM-Agent / Asystenci24) · https://twinlix.com
  Multi-tenant AI voice-agent SaaS: inbound call handling, multi-channel (Twilio, Instagram, Facebook), knowledge management with document ingestion (6+ formats) synced to ElevenLabs, **RAG on Qdrant**, Stripe billing.
  *Tech: Node.js/Fastify, TypeScript, MySQL + Prisma, ElevenLabs Conversational API, n8n, **Qdrant**, PM2.*
- **One-One — multi-persona AI companion platform** · https://one-one.online
  Multi-provider LLM engine: per-slot model routing with automatic fallback (OpenAI-compatible, DeepSeek, self-hosted Ollama/vLLM), dedicated prompt-builder module, persistent memory, and a voice fast-lane with instrumented latency/token benchmarking and per-call model-override for A/B comparison.
  Tech: Python/FastAPI, vLLM, Ollama, DeepSeek.

- **Ordora — CRM assistant with native RAG memory**
  Hand-built vector-retrieval memory from scratch: embeddings (embeddinggemma, 768-dim), cosine similarity, Float32 vector store, a `search_memory` tool doing top-k / min-score retrieval and injecting past context into an agentic tool-calling loop.
  Tech: Node.js, OpenAI-compatible embeddings, custom vector store.
- **Document OCR & entity extraction** — vision-LLM service
  `/ocr` (verbatim extraction) and `/receipt` (**structured JSON** from receipt/invoice photos — key entities + attributes), with field validation.
  Tech: Python/FastAPI, Google Gemini Vision.
- **Real-time voice agents** (Callyou / Hermes)
  Browser and telephony voice loops: Whisper STT → LLM → Silero/Piper/ElevenLabs TTS over WebSockets.
  Tech: Python/FastAPI, Whisper, Silero, Piper, ElevenLabs.
- **Lanehub — Telegram hub for people + AI agents**
  Self-hosted bridge routing conversations between people and AI agents.
  Tech: Python, Docker.

Code & repos: https://github.com/grabbly

Certifications

Anthropic Education — completed courses (independently verifiable):

- **Claude 101** — Aug 2026 — https://verify.skilljar.com/c/ro8btuhi7vvo
- **Building with the Claude API** — Aug 2026 — https://verify.skilljar.com/c/sekkj7zuoba9
- **Claude Code in Action** — Aug 2026 — https://verify.skilljar.com/c/yuvti3ukdb3s
- **Introduction to Subagents** — Aug 2026 — https://verify.skilljar.com/c/rhcx7cevpxrf
- **Introduction to Agent Skills** — Jul 2026 — https://verify.skilljar.com/c/239vew39hyjh

Education

- **Bachelor’s degree** — Software Engineering / Automated Systems
  Mykolaiv State Shipbuilding University (UDMTU), Ukraine

Languages

- English — B1
- Ukrainian — Native · Russian — Native
- Polish — B1

What I’m Looking For

An **AI Engineer** role building and shipping LLM-backed systems — RAG pipelines, model integration and wrappers, voice and document AI, and evaluation-driven iteration — on a cross-functional team that values production reliability and pragmatic mixing of ready-made and custom solutions.

I agree to the processing of my personal data contained in this CV for the needs of the recruitment process, in accordance with Regulation (EU) 2016/679 (GDPR).