Vladislav Vitovtsev

AI Engineer (Agentic Systems, RAG, Fullstack)

Email: vlad@vitovtsev.site **Telegram:** @vladislav_vitovtsev **Site:** vlad.vitovtsev.site **Based in Montenegro** (CET), remote
Contract: my Montenegrin company (d.o.o.), invoices in EUR

SUMMARY

AI engineer with 13 years of fullstack development (Node.js, Python, React). For the past year I have been building agentic AI systems and wiring them into a team's development process. I designed and shipped a RAG service that cut the context sent to the model from 43 KB to 5 KB per query at recall@5 = 1.00. I built agentic code review that, on a single feature, sent back 11 of 17 tasks with real defects before merge: an open redirect, race conditions.

I take to production what usually gets stuck at the demo stage: tool failure handling, escalation to a human, memory between runs, context cost. I own products from idea to running service: backend, frontend, CI/CD.

Open to project work and full-time roles. Async by default: specs and reviews in writing.

EXPERIENCE

Product IT company · video communication product, under NDA — Sep 2025 - present

Sole frontend developer

- Grew a demo prototype of a few dozen files into a production web client: WebRTC calls with recording, calendar with recurring meetings, chat, SSE notifications, virtual background, screen sharing.
- Shipped an Electron desktop client with channel-based auto-updates.
- Built CI/CD for macOS and Windows with code signing (Developer ID with notarization, hardware token via PKCS#11): **49 releases in 7 months**, 1-2 per week on average. Moved the project to a monorepo, about 3,500 tests.
- Set up an agentic development pipeline: spec, plan, test-first, review by separate agents per domain (security, correctness, performance). On an 18-task feature, reviewers sent back **11 of 17** tasks with real defects: an open redirect, two race conditions on the unread counter, a subscription leak. All fixed before merge.
- An agent investigates incidents from Grafana logs and metrics and drafts a ticket with the root cause and fix options; it goes to the tracker after my approval.

Independent practice · consulting, alongside main job — Nov 2023 - present

Lead developer / AI engineer: technical director, sole developer, team developer

Several client projects in parallel: products built from scratch and AI integration.

- **SEO platform** (2026, AI engineer): AI content analysis module for the client's site. Dockerized service with an API, local embedding model, claim-based LLM judge, competitor discovery from Yandex search results. Moved the judge from eyeballed scores to claim checks that show which claims competitors already cover. Switching the judge to a faster model cut a full check from 16.5 s to 5.6 s. An audit closed SSRF, a concurrency-limit race and CSV injection; 145 tests. As a second service, built a page-quality rater based on search quality rater methodology.
- **Caerus** (2024-2026, sole developer): the site of a car rental owned by one of GlobalRental's founders. In 2024 I launched it on the GlobalRental engine as a second brand, in 2025 split the frontends, and in 2026 separated it completely: both sites ran one binary on one database with no multi-tenancy, so first an audit and a rollback point, then its own database and backend, with bookings and the dashboard moved over without losses. Since then, bug fixes and support. The dashboard became the rental's working calendar (about 1,400 entries over two years); direct bookings from the site went from 13 in 2025 to 31 in 2026: **caerusrentacar.com**
- **Psychotherapy platform** (2025-2026, frontend in a team): the whole "Supervision" feature, from supervisor search to payment, rescheduling and calendar. Next.js 15, Nx, Feature-Sliced Design. Delivered it with my agentic process: 4-5 hours against my own manual estimate of 19.
- **EVIRMA** (2025, frontend in a team of 7-8): Manifest V3 Chrome extension with ad analytics for Wildberries sellers. Sales funnel, campaign stats with ad-cost ratios, Excel export. Second-largest contributor after the tech lead.
- **GlobalRental** (2023-2025, technical director): a car rental aggregator in Montenegro. In 2023 I managed the contractor studio on the investor's side: specs, acceptance, production and infrastructure. In 2024 I wrote code alongside the studio's developer: emails, a Telegram bot, an admin panel with stats. From December 2024 I was the only developer on frontend and backend: moderation, currencies, SEO, yacht rentals. **331 of 645** frontend commits, **136 of 294** backend commits. Stripe was verified with a live payment, but the project was shut down before customers started paying online.

App In The Air · USA, travel app with 6M+ users — 2019 - 2024

Frontend developer → team lead of 5

- Joined when the company had a mobile app but no web product, and built the web from scratch: flight search and booking over socket.io, then hotels and a white-label hotel widget for airlines.
- Assembled the frontend team: hiring, onboarding, growth. Ran the technical interviews and take-home reviews myself.

- Built the first version of the corporate B2B portal with one colleague, then led its frontend: employees, corporate cards, booking within travel policy, spend reports.
- Moved the frontend to an Nx monorepo, upgraded React 17 → 18, introduced Next.js, set up a separate Cloud Run environment for every release branch.

webrise · web agency — Aug 2016 - Nov 2019

Fullstack developer

- Redesign of the Sberbank Insurance site and about fifteen insurance calculators with tested pricing logic, calculators for Ingosstrakh.
- A national contest platform with in-browser video and chat, an online store for the Usachevsky market with React Native apps.

Zaymer (microfinance) — Apr 2014 - Aug 2016

Fullstack developer

- Fully automated microloans to bank cards, Russia and Kazakhstan. Laravel 5, MySQL, PostgreSQL, MongoDB, Selenium.

it-infinity — Aug 2013 - Mar 2014

Frontend developer, corporate sites on 1C-Bitrix

OWN AI SYSTEMS AND PROJECTS

- **RAG service in production**, Python and FastAPI: hybrid search (embeddings plus BM25 via SQLite FTS5, merged with RRF), no separate vector DB. Context to the model dropped from 43 KB to 5 KB per question; on 45 control questions recall@5 = 1.00, p95 search under 350 ms. If search fails, the request falls back to the full knowledge base and the visitor never sees the failure.
- **vlad.vitovtsev.site**: portfolio with a live AI agent at the entrance. Fastify, SSE streaming, spend cap, prompt-injection protection, per-IP limits.
- **Agent memory** across sessions: daily snapshots, decision log, auto-checkpoints, my own search over 20 thousand chunks. Cut startup context from 21,700 to 12,000 tokens per run. 23 of my own extensions for the agent environment, a Yandex API toolset built on open-source skills.
- **Helm**: a phone console for parallel agent sessions. Svelte 5, in-browser terminal, PWA, passkey login, about 1,200 tests.
- **trainerpad.ru**: offline-first CRM for personal trainers, solo from scratch. Custom IndexedDB sync engine with conflict resolution.

STACK

AI	agentic systems, RAG, hybrid search, embeddings, prompt engineering, LLM APIs (Anthropic, OpenAI-compatible), MCP, token cost control
Backend	Node.js, Python, FastAPI, Fastify, Express, PostgreSQL, SQLite, Prisma
Frontend	TypeScript, React, Next.js, Electron, React Native, Flutter, WebRTC, SSE, Vitest, Playwright
Infra	Docker, GitLab CI, GitHub Actions, Linux, Nginx, Caddy, own VPS

CERTIFICATES

- **GREEN-API Qualification Certificate** No. 2243758, September 2026: integrations and chatbot development on GREEN-API (a gateway to WhatsApp, MAX and Telegram). Hands-on part: a web chat for MAX on their HTTP API, React 19, TypeScript, Zustand, notification queue, delivery statuses, 385 tests.
  Demo: max-chat.vitovtsev.site

EDUCATION AND LANGUAGES

Tomsk State University of Control Systems and Radioelectronics, 2012, computer engineering. Russian native. English: fluent in writing, which is how I run projects.