

Vladislav Vitovtsev

AI Engineer · Fullstack Developer · 13 years of experience

Email: vlad@vitovtsev.site Site: vlad.vitovtsev.site Montenegro (CET), remote

SUMMARY

Thirteen years of building products end to end: interface, backend, build pipeline, rollout. For the past two and a half years I have been running two projects in parallel. For the past year I have been building agent-based AI systems and putting them into a working team's process, as part of development rather than as an experiment.

I ship the parts that usually stall at the demo stage: tool failure handling, escalation to a human, memory between runs, context cost.

Available for contract work, async by default: specs and reviews in writing, no standing calls.

AGENT-BASED AI SYSTEMS

- **RAG service in production** (July 2026), Python and FastAPI: a couple of hundred chunks, hybrid retrieval with cosine similarity over numpy embeddings and BM25 over SQLite FTS5, merged with RRF. Context sent to the model dropped from 43 KB to 5 KB per question. On a 45-question golden set: recall@5 of 1.00, p95 search under 350 ms, full live answer 1.7 s. A search timeout falls back to the full knowledge base, so a failure of the service is invisible to the visitor.
- Subagent orchestration: task breakdown, spec, plan, TDD execution, review by several agents with their own domains (security, correctness, performance), visual verification on a staging build. Custom tools for agents and integrations over MCP. On an 18-task feature, reviewers returned more than half of the tasks with real defects: an open redirect, two counter race conditions, a subscription leak.
- Wrote 23 extensions for the agent environment. Among them a set for Yandex APIs (Metrica, Wordstat, Webmaster, Search API) with caching and quota handling, since every ready-made tool in this area targets Google.
- Memory system across sessions: daily and weekly digests, decision log, automatic checkpoints, sharding for parallel sessions. Cut the cost of the startup context from 21,700 to 12,000 tokens per launch without losing answer quality.
- **vlad.vitovtsev.site** - a portfolio where a live agent answers questions about my work at the entrance: Node and Fastify, streaming over SSE, curated knowledge base, intent detection that forwards a contact to my Telegram, a hard spending cap, protection against prompt injection, per-IP limits. Built in two days, 226 tests.

EXPERIENCE

Product IT company · under NDA

September 2025 - present

Sole frontend developer on the project

- Joined when the project was a demo prototype of about twenty files and grew it into a production web client: WebRTC calls with recording, calendar with recurring meetings, chat, notification feed over SSE, virtual background, screen sharing.
- Electron desktop client on top of the web app, with channel-based auto-updates.
- Moved the project to a monorepo, set up the release process and CI with macOS and Windows builds: Developer ID signing with notarization, Windows signing through a hardware token over PKCS11.
- **34 release tags in five months**, 758 test cases across 116 files.

Caerus · project work alongside the main job

March 2026 - July 2026

Sole developer

- Separated a partner's site from a shared car rental aggregator. In production both sites ran the same binary against the same database and differed only by environment variables, with no multi-tenancy in the schema: I audited production, took a snapshot as a rollback point, and then split them.
- Have run the product alone since: SEO infrastructure for organic traffic, human-readable URLs, production fixes from partner reports.
- The site is live and earns money for the client: **caerusrentacar.com**

Psychotherapy and coaching platform · project work alongside the main job

June 2025 - April 2026

Frontend developer in a team

- My area was the Supervisions feature end to end: supervisor profile, public search with filters, purchase by a client, specialist dashboard with payment methods, session cancellation and rescheduling, calendar integration.
- Next.js 15 App Router in an Nx monorepo, Feature-Sliced Design, Zustand, CSS Modules.
- Ran the feature through an agent-based spec-driven process: 4-5 hours of work against my own manual estimate of 19 hours.

EVIRMA · project work alongside the main job

February 2025 - September 2025

Frontend developer in a team of 7-8

- Manifest V3 Chrome extension for Wildberries marketplace sellers: embeds advertising analytics directly into the marketplace interface. Intercepting fetch inside someone else's markup, token bucket rate limiting, background jobs in a service worker

that survive suspension.

- Built the seller sales funnel from scratch, campaign statistics by keyword and day with ad-cost ratios, export into multi-sheet Excel workbooks, charts.
- Second by delivered scope in the team, after the tech lead. evirma.ru

Car rental aggregator · project work alongside the main job

March 2024 - August 2025

Frontend developer

- Several rental companies upload cars, a customer searches by city and dates and books. My areas: catalogue with faceted filters and price calculation, booking with availability checks against a calendar, partner dashboard with a booking timeline, admin panel with roles, six languages, payments through Stripe and PayPal.
- Worked through pull requests from tickets, with code review on every change.

App In The Air · USA

November 2019 - October 2024

Frontend developer, then lead of a team of 5

- Joined when the company had a mobile app and no web. Built flight search and ticket purchase: fare selection, results, passenger data, payment. Then hotel booking.
- Assembled the team, grew it to five people, and ran both the team and the product.
- Took the product to a corporate B2B portal: employees, corporate cards, booking within a travel policy, expense analysis.
- Travel provider and payment integrations went through the company's own backend; on the frontend, Stripe, i18next with ICU, Sentry.
- Moved the frontend to an Nx monorepo, upgraded React 17 to 18, introduced Next.js.
- Worked until the service shut down in October 2024.

webrise · web agency

August 2016 - November 2019

Fullstack developer

- Full redesign of the Sberbank Insurance site and around fifteen insurance calculators, including internal ones for bank staff, plus a calculator line for Ingosstrakh. A mistake in the maths meant a wrong policy price for a customer, so the calculation layer was covered by tests.
- Learning platform for a nationwide competition with video and in-browser chat (bolshayaperemena.team), online store for Usachevsky Market with React Native mobile apps (usch.ru).
- Promo projects for Unilever (Calve, Baltimor, Clear, Rexona), a game for Ritter Sport, a digital platform for a transport university.

Zaymer

April 2014 - August 2016

Fullstack developer

- Microloan service paying out to cards with no human in the loop, Russia and Kazakhstan: customer and staff dashboards. Laravel 5, MySQL, PostgreSQL, MongoDB, Selenium tests.

it-infinity

August 2013 - March 2014

Frontend developer - redesign and markup of corporate sites on top of 1C-Bitrix

STACK

Languages	TypeScript, JavaScript, Python, PHP (in the past)
Frontend	React, Next.js, Preact, Astro, Electron, WebRTC, SSE and WebSocket, Vitest, Playwright
Backend	Node.js, Fastify, Express, FastAPI, PostgreSQL, SQLite, Prisma
AI	agent orchestration, MCP, prompt engineering, LLM APIs (Anthropic, OpenAI-compatible), RAG with hybrid retrieval and embeddings, token cost control
Infrastructure	Docker, GitLab CI, GitHub Actions, Caddy, Nginx, Linux and my own VPS, desktop build signing and notarization

OWN PROJECTS

- **vlad.vitovtsev.site** - a portfolio built as an engineer's desktop: window interface on plain DOM and CSS, Astro with a Preact island, AI agent at the entrance. First screen 25 KB, all JS under 20 KB.
- **trainerpad.ru** - offline-first CRM for personal trainers, built solo from scratch: 626 commits, custom sync engine on IndexedDB with conflict resolution and cross-tab locking instead of an off-the-shelf solution.

EDUCATION AND LANGUAGES

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

Based in Montenegro (CET). Two ways to contract: my Montenegrin company (d.o.o.) with invoices in EUR, or my sole proprietorship in Russia with payment in RUB.