



Roman Yastrebov

Team Lead / AI Engineer

Saint Petersburg, working remotely Telegram @omasn yastrebov.tech

10 years in backend development with PHP/Symfony, Go and Python, 4 of them as an architect and team lead. I design products from scratch and pull heavy legacy systems back into shape. Since 2025 I have been working with LLMs in products: deploying open-weight models, building LLM gateways, RAG and AI assistants for production. On my latest legacy project I moved the team to agentic development with Claude Code, and tasks started getting done 40% faster.

Key skills

Backend and architecture	PHP 8 / Symfony, Go (gRPC microservices), Python (FastAPI);DDD, Hexagonal, microservices and modular monoliths; asynchronous processing with queues.
AI Engineering	Agent harness (tool calling, streaming, memory), self-hosted open-weight models (vLLM, Ollama), Whisper with diarization, LiteLLM, RAG on pgvector, eval sets, model cascades; OpenAI, Anthropic, Gemini APIs.
Agentic development	Claude Code, Codex, Cursor; spec-driven development, MCP integrations of agents with the task tracker and GitLab, agent code review with test verification.
Engineering standards	Functional tests for the whole API, TypeScript models generated from OpenAPI, static analysis, architectural boundary checks (Deptrac).
Data and operations	PostgreSQL, Elasticsearch, Redis, RabbitMQ, Docker, GitLab CI/CD, Prometheus, Grafana, Graylog, Sentry.

Experience

Code Pilots Custom software studio 2017 – present	Started as a backend developer, then became an architect and team lead; now a team lead and AI engineer.
GigWork Outsourcing January 2023 – December 2025	Team lead of GigWork Exchange, then architect of the HRM platform on PHP/Symfony and Go.

Projects

2025 – present Code Pilots and own products

AI infrastructure and LLM products

- Designed an AI assistant for a Flutter mobile app on a custom agent harness: a tool-calling loop with tools over user data, web search, streamed answers, memory and parsing of uploaded documents. Models are connected through LiteLLM, so the provider changes without code edits and a failed request falls back to a backup model.

- Took the assistant to production: daily quotas, an answer deadline, token accounting and a trust boundary for tool results against prompt injection. Prompts are versioned with the code and updated without an app release.
- Deployed an LLM gateway on LiteLLM: one OpenAI-compatible API for cloud and self-hosted models, per-project keys and budgets, fallbacks between providers.
- Ran open-weight models on my own servers for tasks where data cannot go to external APIs: LLMs on vLLM and Ollama, speech recognition with Whisper and meeting diarization.
- Built RAG search over documents on PostgreSQL with pgvector, with hybrid search and an eval set of questions I use to check answer quality.

April 2026 — present Code Pilots Team lead and AI engineer underdogi.ru

Underdogi

A system that recognizes PDF protocols of sports competitions with LLMs and computes ratings of Kyokushin karate fighters.

- Lead the project team: assign tasks, review code, choose models and how recognition quality is measured.
- Built an eval bench and used it to pick models for recognizing 30k pages of protocols. Accuracy went from 53% to 96% at about \$0.01 per page.
- Designed a new recognition engine. The model extracts the bracket structure and code rebuilds the tournament logic; disputed pages go to a stronger model, and spending is capped.
- Sped up the rating page from 10 seconds to 60 ms with precomputed snapshots.
- Added validation and deduplication of results, a data snapshot and automatic rollback on failure to the weekly run.

Stack: Python, FastAPI, OpenAI API, Gemini API, PHP 8.4, Symfony 7.4, MySQL, Nuxt 3, Docker.

January 2024 — December 2025 GigWork, outsourcing Architect gigwork.ru

GigWork

An HRM platform for staff outsourcing. Several PHP/Symfony services and Go microservices.

- Owned the platform architecture. Designed new modules, agreed technical decisions with the team and introduced architecture reviews for major changes.
- Wrote Go microservices for the high-load parts of the platform: RabbitMQ event handlers and services with gRPC APIs.
- Moved mass accruals and payouts to contractors to asynchronous batch processing so that peaks do not block the API.
- Designed piece-rate pay across the whole process: order, assigning contractors, reconciliation, payouts, reporting.
- Proposed and carried out merging the customer portal and HRM into one service. Duplicated code went away and new features started shipping about 25% faster.
- Upgraded the services from PHP 7.4 and Symfony 5.4 to PHP 8.3 and Symfony 6.4 and covered the whole customer portal API with functional tests.

Stack: Go, gRPC, PHP 8.3, Symfony 6.4, PostgreSQL, Redis, RabbitMQ, Graylog, GitLab CI/CD.

January – April 2026 Code Pilots

Go-Base Platform

An ERP/WMS warehouse system with RFID. Heavy legacy: PHP 7.0, a custom framework on Zend Framework 1, 58 business modules.

- Moved the team to agentic development with Claude Code; tasks started getting done 40% faster.
- Introduced a spec-driven process and automated the routine around tasks: agents create a branch from a YouTrack task, open a merge request and update the status on their own.
- Migrated the codebase from Zend Framework 1 to Symfony components with no regressions for neighbouring teams.

Stack: PHP, Zend Framework 1, Symfony, MariaDB, PostgreSQL, RabbitMQ, gRPC, Claude Code, MCP.

January 2023 – January 2024 GigWork, outsourcing Team lead market.gigwork.ru

GigWork Exchange

A request exchange for outsourcing companies in the CIS. A product built from scratch.

- Led a team of 8: 4 backend, 2 frontend, 2 QA.
- Laid down the architecture on DDD and Hexagonal, with module boundaries checked by Deptrac.
- Set up TypeScript model generation from OpenAPI: the frontend received contracts automatically, without manual syncing.

Stack: PHP, Symfony, PostgreSQL, Deptrac, Psalm, PHPStan, OpenAPI, GitLab CI/CD.

September 2021 – January 2023 Code Pilots Architect, then team lead toomanygifts.ru

Too Many Gifts

A B2B store of corporate gifts with over a million products in the catalogue. A team of 4.

- Designed the platform from scratch: the main backend and an import microservice connected through RabbitMQ.
- Built faceted catalogue search on Elasticsearch; filters and counters are computed with aggregations without database queries.
- Designed the import from 15+ suppliers with different protocols (REST, SOAP, FTP, XML feeds) and gigabytes of data. Parsers stream within bounded memory, synchronization is incremental, and a new supplier is added with a single parser.
- Relunched the live store on the new platform without losing SEO.

Stack: PHP 8.1, Symfony 5.4, PostgreSQL, Elasticsearch, Redis, RabbitMQ, S3, Docker.

Earlier projects

Code Pilots, 2019 – 2022

Blaze

A platform for the beauty industry: online booking, room rental, subscriptions. Architect and lead backend developer. A backend of 15 isolated domain modules, billing with subscriptions and acquiring, integrations with MoySklad and RetailCRM, metrics and alerts on Prometheus and Grafana.

MiroFS

A fitness app for iOS and Android. Wrote the whole backend: an API for both apps, an admin panel for trainers, subscriptions via Apple In-App Purchase with server-side receipt validation, Yookassa and CloudPayments.

Radio Record

A radio station mobile app. An API for the app on top of the existing website with logic on Symfony components, paid subscriptions with recurring payments, and a Node.js service that extracts the current track and cover art from the radio stream in real time.

CV updated in September 2026.