

Роман Ястребов

Team Lead / AI Engineer

Санкт-Петербург, работаю удалённо Telegram @omasn yastrebov.tech

10 лет в backend-разработке на PHP/Symfony, Go и Python, из них 4 года архитектором и тимлидом. Проектирую продукты с нуля и вытаскиваю тяжёлое легаси. С 2025 года работаю с LLM в продуктах: разворачиваю модели с открытыми весами, строю LLM-шлюзы, RAG и AI-ассистентов для продакшена. На последнем легаси-проекте перевёл команду на агентную разработку с Claude Code, и задачи стали закрываться на 40% быстрее.

Ключевые компетенции

Backend и архитектура	PHP 8 / Symfony, Go (gRPC-микросервисы), Python (FastAPI); DDD, Hexagonal, микросервисы и модульные монолиты; асинхронная обработка через очереди.
AI Engineering	Agent harness (tool-calling, стриминг, память), self-hosted модели с открытыми весами (vLLM, Ollama), Whisper с диаризацией, LiteLLM, RAG на pgvector, eval-наборы, каскады моделей; OpenAI, Anthropic, Gemini API.
Агентная разработка	Claude Code, Codex, Cursor; spec-driven development, MCP-интеграции агентов с трекером и GitLab, ревью кода агентами с проверкой тестов.
Инженерные стандарты	Функциональные тесты на весь API, генерация TypeScript-моделей из OpenAPI, статический анализ, контроль архитектурных границ (Deptrac).
Данные и эксплуатация	PostgreSQL, Elasticsearch, Redis, RabbitMQ, Docker, GitLab CI/CD, Prometheus, Grafana, Graylog, Sentry.

Опыт работы

Code Pilots Студия заказной разработки 2017 — настоящее время	Начинал backend-разработчиком, потом стал архитектором и тимлидом, сейчас тимлид и AI-инженер.
GigWork Аутсорс Январь 2023 — декабрь 2025	Тимлид «GigWork Биржи», затем архитектор HRM-платформы на PHP/Symfony и Go.

Проекты

2025 — настоящее время Code Pilots и свои продукты

AI-инфраструктура и LLM-продукты

- Спроектировал AI-ассистента для мобильного приложения на Flutter на собственном agent harness: цикл tool-calling с инструментами по данным пользователя, веб-поиск, стриминг ответов, память, разбор загруженных документов. Модели подключены через LiteLLM, поэтому провайдер меняется без правок кода, а при сбое запрос уходит на запасную модель.

- Довёл ассистента до продакшена: дневные квоты, дедлайн на ответ, учёт токенов, граница доверия для результатов инструментов против prompt injection. Промпты версионизируются вместе с кодом и обновляются без релиза приложения.
- Развернул LLM-шлюз на LiteLLM: один OpenAI-совместимый API для облачных и своих моделей, ключи и бюджеты по проектам, фолбэки между провайдерами.
- Поднял модели с открытыми весами на своих серверах для задач, где данные нельзя отдавать во внешние API: LLM на vLLM и Ollama, распознавание речи на Whisper и диаризация встреч.
- Сделал RAG-поиск по документам на PostgreSQL с pgvector, с гибридным поиском и eval-набором вопросов, по которому проверяю качество ответов.

Апрель 2026 — настоящее время Code Pilots Тимлид и AI-инженер underdogi.ru

Underdogi

Система автоматического распознавания PDF-протоколов спортивных соревнований через LLM и расчёта рейтингов бойцов кёкусинкай-карате.

- Веду команду проекта: распределяю задачи, ревьюю код, выбираю модели и методику оценки качества распознавания.
- Собрал eval-стенд и по нему подобрал модели для распознавания 30 тыс. страниц протоколов. Точность выросла с 53% до 96% при стоимости около \$0,01 за страницу.
- Спроектировал новый движок распознавания. Модель извлекает структуру сетки, а логику турнира восстанавливает код; спорные страницы уходят на более сильную модель, на расходы стоит лимит.
- Ускорил страницу рейтинга с 10 секунд до 60 мс за счёт предрассчитанных снимков.
- Добавил в еженедельный прогон валидацию и дедупликацию результатов, снимок данных и автоматический откат при сбое.

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

Январь 2024 — декабрь 2025 GigWork, аутсорс Архитектор gigwork.ru

GigWork

HRM-платформа для аутсорсинга персонала. Несколько сервисов на PHP/Symfony и микросервисы на Go.

- Отвечал за архитектуру платформы. Проектировал новые модули, согласовывал технические решения команды и ввёл архитектурное ревью для крупных изменений.
- Писал микросервисы на Go для нагруженных частей платформы: обработчики событий из RabbitMQ и сервисы с gRPC-API.
- Перевёл массовые начисления и выплаты исполнителям на пакетную асинхронную обработку, чтобы пики не блокировали API.
- Спроектировал сдельную оплату через весь процесс: заказ, назначение исполнителей, сверка, выплаты, отчётность.
- Предложил и провёл объединение личного кабинета и HRM в один сервис. Дублирование кода ушло, новые функции стали выходить примерно на 25% быстрее.
- Обновил сервисы с PHP 7.4 и Symfony 5.4 до PHP 8.3 и Symfony 6.4 и покрыл API личного кабинета функциональными тестами целиком.

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

Go-Base Platform

ERP/WMS-система складского учёта с RFID. Тяжёлое легаси: PHP 7.0, самописный фреймворк на Zend Framework 1, 58 бизнес-модулей.

- Перевёл команду на агентную разработку с Claude Code, задачи стали закрываться на 40% быстрее.
- Ввёл спес-driven процесс и автоматизировал рутину вокруг задач: агенты сами заводят ветку по задаче из YouTrack, открывают Merge Request и обновляют статус.
- Переводил кодовую базу с Zend Framework 1 на компоненты Symfony без регрессий у смежных команд.

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

GigWork Биржа

Биржа заявок для аутсорсинговых компаний в СНГ. Продукт с нуля.

- Руководил командой из 8 человек: 4 backend, 2 frontend, 2 QA.
- Заложил архитектуру на DDD и Hexagonal, границы модулей проверял Deptrac.
- Настроил генерацию TypeScript-моделей из OpenAPI: фронтенд получал контракты автоматически, без ручной синхронизации.

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

Too Many Gifts

B2B-магазин корпоративных подарков, в каталоге больше миллиона товаров. Команда из 4 человек.

- Спроектировал платформу с нуля: основной бэкенд и микросервис импорта, между ними RabbitMQ.
- Сделал фасетный поиск по каталогу на Elasticsearch, фильтры и счётчики считаются агрегациями без запросов в БД.
- Спроектировал импорт от 15+ поставщиков с разными протоколами (REST, SOAP, FTP, XML-фиды) и гигабайтами данных. Парсеры потоковые и укладываются в ограниченную память, синхронизация инкрементальная, новый поставщик подключается одним парсером.
- Перезапустил работающий магазин на новой платформе без потери SEO.

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

Ранние проекты

Code Pilots, 2019 — 2022

Blaze

Платформа для бьюти-индустрии: онлайн-запись, аренда кабинетов, подписки. Был архитектором и ведущим backend-разработчиком. Бэкенд из 15 изолированных доменных модулей, биллинг с подписками и эквайрингом, интеграции с МойСклад и RetailCRM, метрики и алерты на Prometheus и Grafana.

MiroFS

Фитнес-приложение для iOS и Android. Написал весь бэкенд: API для двух приложений, админку для тренеров, подписки через Apple In-App Purchase с серверной проверкой чеков, ЮKassa и CloudPayments.

Radio Record

Мобильное приложение радиостанции. API для приложения поверх существующего сайта с логикой на компонентах Symfony, платные подписки с автосписаниями и сервис на Node.js, который в реальном времени достаёт из радиопотока текущий трек и обложку.

Резюме обновлено в сентябре 2026 года.