Феликс Десятириков

AI Engineer, **6 лет** опыта
GMT+3

Telegram: @felix_des | E-mail: felixdesyatirikov@gmail.com | Github.com/FelixDes

Обо мне

AI-разработчик с коммерческим опытом работы **6 лет** в командах от 5 до 20 человек. Интегрировал платформенные ИИ-решения в действующие и проектируемые информационые системы. Выстроил **полный цикл разработки**: от сбора требований до реализации системы с автоматизированной системой релизов с высоким покрытием тестами и использованием **передовых практик CI/CD и DevOps**.

В дополнение к основным знаниям по LLM/AI имею богатый опыт работы с Kubernetes и микросервисной архитектурой. Интересуюсь развитием экосистемы PyTorch, постоянно обновляю свои знания по этим темам. Мое хобби — развитие личного бренда через доклады на профильных конференциях, публикацию научных статей и вклад в Open Source.

Навыки

AI/ML	Python, FastAPI, LangChain, LangGraph, RAG, NLP, PyTorch, FAISS, Qdrant, PGvector, Bert, MLops, DevOps, TTS, STT, Hugging Face, asyncio, CatBoost, XGboost, LGBM, LoRA, QLoRA, oLLaMA, llama.cpp, RL, RLHF, SFTJava, Kotlin, mlfow, PostgreSQL
LLM models	Llama, Mistral, OpenAi ChatGPT, DeepSeek, Anthropic, Grok
Big Data	Apache Kafka, Hadoop, Spark, Hive, Impala, Airflow
DevOps	Kubernetes, OpenShift, Microservices, Docker, Vagrant, Docker Compose, CI/CD, Ansible, Linux, Prometheus, Grafana, S3, MinIO, Google Cloud Platform
Architecture	DDD, TDD, BDD, Hexagonal Architecture, Microservices

Опыт работы — 6 лет

VK

6 месяцев
Июль 2025 - н/в

Ведущий AI Инженер

- Реализовал **RAG-систему** для анализа инцидентов и **автоматического обновления документации SRE**. Это позволило **сократить MTTR инцидентов на 23%** и ускорить подготовку новых SRE инженеров на **10%**
- Интегрировал сбор метрик и трассировок в **MCP-based систему**. Это повысило observability, позволило добиться **повышения SLA с 99.2% до 99.85%**, улучшить прогнозируемость сбоев и ускорить локализацию инцидетов
- Внедрил в **MCP систему** технической поддержки сотрудников модуль интеграции с in-house OLAP хранилищем. Это **позволило автоматизировать типовые сценарии выдачи доступов** и успешно решать **37%** проблем по данном домену в автоматическом режиме
- Отвечаю за техническое развитие команды: провожу тех. интервью, помогаю в онбордингах и организую внутренние воркшопы

Stack: Python, RAG, LLM, FastAPI, LangChain, Java, Spring Boot

T1 Innotech

1 год
Июль 2024 - Июль 2025

Ведущий AI Инженер

- Провел полный цикл **проектирования, реализации и внедрения** сервиса **валидатора клиентского чат-бота** на основе **Mistral 3 8B** и **LoRA**. Внедрение сервиса позволило обеспечить безопасность ответов системы
- Построил **многоагентную MCP** систему навигации по документации разных бизнес-доменов. Она позволила облегчить процесс написание документации и уменьшила срок онбординга новых сотрудников на **16%**
- Реализовал подсистему **AI ассистента** для ответа на письма в in-house почтовом клиенте с учетом истории сообщений и использования персонализированных мастер промтов пользователей. AI-ассистент помогал обрабатывать **до 28% входящих писем автоматически**

Stack: Python, FastAPI, LLM, Mistral, LoRA, LangChain, Lang Graph, RAG, MCP, PGvector, Apache Kafka

Bell Integrator

1 год 8 месяцев
Нояб. 2022 - Июль 2024

AI Инженер

- Добавил в **модуль валидатора RAG-системы** сценарий, выявляющий запросы с недостаточными правами доступа и предлагающий пользователю повышение привилегий, что устранило ложные ответы об отсутствии данных и **повысило прозрачность системы**
- Реализовал систему **семантичесского кеширования** запросов в **RAG** систему. Это улучшение позволило получить **cache hit-rate до 90%** на семантически близких запросах, **понизить latency на 13.5%**, **сократить LLM-inference расходы на 12%** и повысить **удовлетворенность пользователей**
- Релизовал систему уведомлений о нахождении персональных данных в базе знаний, что повысило уровень безопасности исходного источника данных

Stack: Python, FastAPI, LangChain, RAG, NLP, PGvector, Apache Kafka

Inline Group

1 год 8 месяцев
Март 2021 - Нояб. 2022

ML-ops/Backend Разработчик

- Разработал **Computer Vision** сервис с использованием YOLO для сокращения времени анализа дефектов на **20.7%**
- Разработал сервис просмотра аналитики результатов действий работника, повысив скорость анализа workflow на **30%**

Stack: Python, FastAPI, OpenCV, YOLO, Docker, PostgreSQL, CI/CD, Java, Spring Boot

Стройсервис

9 месяцев
Июнь 2020 - Март 2021

Full-stack Разработчик

Stack: Python, FastAPI, SQLAlchemy, PostgreSQL, React, Java

Образование

«Прикладная математика и информатика» - Воронежский государственный университет, Воронеж, Россия
Магистратура

«Прикладная математика и информатика» - Воронежский государственный университет, Воронеж, Россия
Бакалавриат

Сертификаты

Architecting with Google Kubernetes Engine Specialization — Март 2024

Публикации

Scopus | Web of Science | IEEE*Xplore*

Знание языков

Русский Родной

Английский Professional working proficiency