

DMITRY ZOLOTAREV

AI / ML Engineer

 dimkabin.github.io

 dmitryellison@gmail.com

 +7 953 010 09 33

 github.com/dimkabin

 Saint Petersburg

 linkedin.com/in/dimkabin

PROFILE

AI / ML Engineer with 3+ years of experience building production LLM, RAG, and agentic AI systems, as well as computer vision services and ML backends. Strong in model integration, evaluation, APIs, observability, and deployment. Built a LangGraph-based harness for an AI analyst with P50 latency of 24s and P99 latency of 51s.

SKILLS

ML / LLM: PyTorch, Transformers, LlamaIndex, LangChain, LangGraph, RAG, MCP, DeepEval, OpenCV, scikit-learn, NumPy

Backend: Python, FastAPI, PostgreSQL, Redis, DuckDB, pandas, SSE, JWT

Platform: Linux, Docker, Kubernetes, Kafka, AWS S3, CI/CD, Nginx, Phoenix, OpenTelemetry

PROFESSIONAL EXPERIENCE

02/2025 – Present	ML Engineer LLM / Agents Innotech	- Built production RAG and document-vectorization pipelines processing 10,000+ documents using LlamaIndex, Kafka, and AWS S3, including TableRAG-inspired retrieval for structured data. - Built a RAG evaluation service using DeepEval and LlamaIndex evaluation tooling. - Developed MCP servers for RAG, web, and database retrieval with visualization workflows, as well as a deep-research service for iterative planning, tool use, and answer synthesis. - Refactored the LLM data-analyst platform into an agent-centric LangGraph loop, cutting P50 answer latency from 50 seconds to 3 seconds. - Implemented 11 agent tools spanning data analysis, visualization, forecasting, anomaly detection, web search, RAG, memory, planning, and review using pandas, DuckDB SQL, and Plotly, with timeouts and LRU caching. - Delivered an AST-validated Python sandbox, FastAPI-to-React SSE streaming, Parquet and JSON session storage, JWT authentication, and LLM tracing with Phoenix and OpenTelemetry. Stack: LlamaIndex, LangChain, LangGraph, MCP, Kafka, AWS S3, DeepEval, Redis, FastAPI, PostgreSQL, Docker, Kubernetes, DuckDB, Phoenix.
09/2023 – 02/2025	ML Engineer AI Laboratory ITMO University	- Researched video-generation methods for the analysis of homogeneous objects and multi-object tracking. - Built computer-vision and NLP experiments using PyTorch, Transformers, PyTorch Lightning, OpenCV, MMDetection, YOLOv8, and scikit-learn.
03/2024 – 09/2024	ML Engineer TatITneft	- Trained YOLOv8 for passport-field detection, prepared OCR datasets with Label Studio, and fine-tuned EasyOCR, reducing character error rate from 0.47 to 0.31. - Built and deployed a FastAPI OCR backend with a Gradio interface and Docker Compose. - Built an audio-to-structured-data pipeline using whisper.cpp, llama.cpp, JSON extraction, Streamlit, and Telegram bot integration. - Built a speech-to-text model pool with FastAPI, selectable Gradio models, per-model Docker images, and GPU server deployment. Stack: Python, FastAPI, PyTorch, Transformers, OpenCV, YOLOv8, EasyOCR, Whisper, llama.cpp, Docker.
09/2023 – 03/2024	Backend Developer TatITneft	- Developed FastAPI services for printed and handwritten OCR workflows in internal applications. - Implemented PostgreSQL storage, authentication, image-upload APIs, asynchronous processing, logging, pytest tests, and WebSocket delivery of OCR results. - Containerized and deployed services with Docker on Linux servers. Stack: Python, FastAPI, PostgreSQL, PyTorch, pytest, Docker, Docker Compose, Linux.
11/2022 – 06/2023	ML Engineer ReflectME	- Developed a virtual try-on pipeline for a smart-mirror startup. - Trained UNet and YOLOv8 segmentation models for human body and clothing regions, as well as PoseNet for pose estimation. - Trained GAN and WGAN-GP models for clothing-conditioned person generation. Stack: Python, PyTorch, OpenCV, YOLOv8, UNet, PoseNet, GAN, WGAN-GP, Albumentations, NumPy.

EDUCATION		
2026 – 2028	ITMO University Artificial Intelligence, AI Talent Hub.	MSc
2021 – 2025	ITMO University Robotics and Artificial Intelligence, Faculty of Control Systems and Robotics. GPA: 4.3/5.0.	BSc
ADDITIONAL TRAINING		
02/2023 – 06/2023	DLS, MIPT Completed an advanced computer-vision track covering image classification, object detection, segmentation, modern neural-network architectures, and practical model development with PyTorch.	Computer Vision
07/2023	AGNI Workshop Built an MVP computer-vision system for the oil and gas industry. Worked on data annotation, preprocessing, augmentation, model selection, training, and ML backend development with PyTorch.	Computer Vision
02/2024 – 06/2024	DLS, MIPT Completed an NLP track covering text preprocessing, embeddings, sequence models, attention, Transformers, language modelling, and practical NLP projects with PyTorch.	NLP
03/2024 – 04/2024	Gazprom Neft Sirius Selected for the Sirius educational programme following a hackathon. Developed a recommendation-system prototype for finding analogous oil wells using geological and production data.	Energy Analytics
07/2024	Advanced Workshop Worked with point-cloud data for autonomous ground robotics, including real-time 3D object detection, classification, and tracking.	3D Robotics
08/2024	AIRI Summer School Developed a multimodal model combining TrOCR, CLIP, and Co-DETR visual encoders with Qwen2-0.5B. Trained a TrOCR adapter and integrated it with the CLIP adapter and the shared language model.	Multimodal AI
LANGUAGES		
English – B2, Russian – Native		