Oleg Zholobov

AI Engineer

olegzholobov.com · tffriend015@gmail.com · Telegram @sdflux

PROFILE

AI Engineer with **5+ years** of experience building ML systems. Focused on LLM agents, language model fine-tuning and enterprise NLP.

EXPERIENCE

viamobi

Jul 2025 - Present

AI Engineer

Conversational LLM agent for a content monetization platform.

- Designed and deployed the agent to production. Versus human operators, it recorded **~30% higher purchase conversion** in chats with sales offers, **84% lower median response time** and a **3.2 pp higher share of returning chats**.
- Built multi-step sales workflows with intent recognition, context-aware responses and adaptive dialogue strategies.
- Fine-tuned a local **Qwen** model with **SFT and GRPO** to improve tool-call sequencing and reduce redundant calls. Applied Schema-Guided Reasoning (SGR), achieving **95% tool-calling accuracy**.
- Built a modular **LangGraph** wrapper to accelerate agent development, with support for graph versioning, nodes, routing and prompts.

CIARS

Apr 2022 - Jul 2025

ML Engineer · ML Team Lead · Head of AI

ML and LLM systems for knowledge retrieval, document processing and intent classification.

- Designed a **GraphRAG** pipeline combining vector search, full-text search and graph-based reasoning; achieved **98% recall and 96% precision** on corporate knowledge-base tests.
- Fine-tuned **Llama / Mistral** with **PEFT and LoRA** using distributed training on an **8-GPU cluster**.
- Built multi-agent LLM pipelines for document classification, summarization and question answering.
- Led teams of **up to three ML and data engineers**, owning technical decisions, mentoring and model deployment.
- Fine-tuned **BERT** for user-intent classification in a marketplace product catalog.

SKILLS

LLM & Agents: RAG · LangGraph · LangChain · OpenAI API · Context engineering · Prompt engineering · MCP · LLM evaluation · LLM-as-a-judge · Langfuse · ElevenLabs · LlamaIndex

Machine Learning: PyTorch · CatBoost · statsmodels · scikit-learn · TRL · Accelerate · JAX · PEFT · LoRA · Unsloth

Backend: Python · SQL · PostgreSQL · pgvector · Qdrant · FAISS · Elasticsearch · ClickHouse · Redis · Airflow

MLOps: Docker · Kubernetes · AWS · GCP · vLLM · GitLab CI/CD · MLflow · Prometheus · Grafana

AI-assisted Development: Claude Code / Codex · Spec-driven implementation and code review

EDUCATION AND LANGUAGES

Financial University under the Government of the Russian Federation 2026

Applied Mathematics and Computer Science

English - C1 · Russian - Native