andrewbelyaev2164@gmail.com
github.com/andreibeliaev
linkedin.com/in/andreibeliaev

Andrei Beliaev

andreibeliaev.github.io

+1 (217) 419 9945
Urbana-Champaign, Illinois
t.me/thelutshe

Summary

AI/ML Software Engineer with 2+ years of experience building ML systems, full stack applications, and data workflows. Developed an agentic platform that lets researchers work with data in natural language, increased automated fraud case resolution from 45% to 60%, improved product quality by 18%, and contributed to $1.7M in annual cost savings.

Experience

Epivara, Inc., BioTech — Champaign, IL
AI/ML Software Engineer — *May 2025 – Present*

- Shipped an agentic data analysis platform with LLM tool-calling that autonomously explores schemas, writes SQL, executes sandboxed Python, and generates reports – orchestrating 10 tools in parallel with persistent state across multi-step reasoning.
- Built a fault-tolerant human-in-the-loop data ingestion agent with guardrails that autonomously parses, validates, and structures experimental datasets, pausing for user approval before writes – reducing insertion time from hours to minutes.
- Adapted foundation vision models for histology analysis by adding task-specific decoder heads and fine-tuning with LoRA on internal and 10k-image public datasets; built the cloud inference backend, moving toward broader use across research labs.
- Customized and deployed a self-hosted annotation platform on Hetzner using Docker and Nuclio, routing GPU inference to Modal for scalable AI-assisted labeling.
- Designed statistical studies that optimized sample sizes, reduced unnecessary costs, and improved reproducibility.

Tinkoff, FinTech — Moscow, Russia
Data Analyst (Compliance) — *Feb 2022 – Oct 2022*

- Contributed to feature development for a new Anti-Money Laundering model that increased fully automated task resolutions from 45% to 60% and cut annual costs by $1.7M.
- Developed SQL/Python/Spark pipelines and dashboards that improved fraud and compliance analysis, cutting 1,000 hours of manual work annually.

Tinkoff, FinTech — Moscow, Russia
Data Analyst (Insurance) — *Mar 2021 – Feb 2022*

- Engineered a real-time quality measurement framework that merged backend logs, frontend events, and database tables, exposing live SRE metrics in Grafana and raising product quality by 18% while saving $60K monthly.
- Collaborated with product, engineering, and support teams to diagnose product issues and run a quality improvement loop.
- Delivered ML-ready datasets for churn prediction and ran A/B tests to measure conversion impacts.

Projects

Entropy-aware sampling in vLLM [Link]

- Implemented entropy-aware token sampling in vLLM with GPU-batched lookahead to control diversity by penalizing entropy-reducing tokens in speculative decoding.

Multi-Model Coding Assistant

- Built a multi-agent coding assistant with OpenClaw integration where a director model plans, challenges, coordinates implementation and review across models, and opens PRs with human review.

Energy-Based Transformers [Link]

- Ran an ablation study of MCMC sampling strategies for Energy Based Transformers in PyTorch Lightning, analyzing pre-training performance across configurations.

Protecting Images from Generative AI Editing [Link]

- Implemented a semantic attack that uses diffusion U-Net cross-attention layers to generate image perturbations, making images more resistant to generative editing at the James M. Rehg Lab (UIUC).

Education

University of Illinois Urbana–Champaign — Champaign, IL
Master's in Computer Science | GPA 3.8 / 4 | GRA — Aug 2024 – May 2026

Bauman Moscow State Technical University — Moscow, Russia
Specialist in Autonomous Informational and Control Systems | GPA 4.7 / 5 — Sep 2016 – Jun 2022

Skills

Languages: Python (FastAPI, SQLAlchemy, asyncio), SQL, TypeScript (React, Zustand, shadcn/ui), Java, C++, MATLAB
AI/ML: PyTorch, Lightning, LangGraph, vLLM, Scikit-learn, W&B, tool and agent orchestration, HITL, RAG
Infra/Data: AWS (EC2, S3, IAM, ECS, VPC), Modal, Docker, PySpark, PostgreSQL, HPC (SLURM)