andrewbeliaev2164@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 a conversational AI 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*

- Fine-tuned SAM-based vision models for histology tissue-scan analysis, achieving 0.94 object-level F1@0.5 IoU and 0.93 pixel IoU on held-out data; trained on the NCSA Delta cluster.
- Deployed and modified a self-hosted CVAT annotation platform on Hetzner (Docker, Nuclio), routing GPU inference to Modal for scalable interactive labeling.
- Shipped a conversational AI analytics app (FastAPI/React, AWS S3) that lets researchers query databases and generate statistical analyses, charts, and reports in natural language, replacing hours of manual SQL and R/Python scripting.
- Built a fault-tolerant HITL data ingestion workflow where an LLM parses, verifies, and structures experimental datasets, reducing insertion time from hours to minutes.
- Delivered statistical study designs that optimized sample size, 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.

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).

Real Time Trade Mirroring System

- Developed a real-time futures trade mirroring service in Python (asyncio, WebSockets) that replicates orders with robust synchronization and failure handling, achieving 56 ms end-to-end latency.

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), C++, MATLAB
ML: PyTorch, Lightning, LangGraph, vLLM, Scikit-learn, W&B
Infra/Data: AWS (EC2, S3, IAM, ECS, VPC), Modal, Docker, PySpark, PostgreSQL, HPC (NCSA Delta, SLURM)