

SUMMARY

**AI/ML Software Engineer** with 2+ years of experience building data workflows, ML tools, and full stack applications. 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**

AI/ML Software Engineer

Champaign, IL

May 2025 – Present

  - (In progress) Adapting SAM-based vision models for tissue-scan segmentation and classification, building AI-assisted annotation tools, and fine-tuning on the NCSA Delta AI cluster.
  - Deployed and modified a self-hosted CVAT annotation platform on Hetzner (Docker, Nucleo), 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**

Data Analyst (Compliance)

Moscow, Russia

Feb 2022 – Oct 2022

  - Contributed to feature development for a new AML 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**

Data Analyst (Insurance)

Moscow, Russia

Mar 2021 – Feb 2022

  - Engineered a real-time quality measurement framework that merged backend logs, frontend events, and database tables, exposing live 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 an iterative 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 an entropy-aware token sampling strategy in vLLM using GPU batched lookahead to control generation diversity by penalizing tokens with high entropy reduction.
  - Logged 22 metrics (including self-BLEU, perplexity, entropy dynamics, acceptance rate) in Weights & Biases for analysis.
- 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 and updates state between accounts with robust synchronization and failure handling, achieving 56 ms end-to-end latency.

EDUCATION

- University of Illinois Urbana-Champaign**

Master's in Computer Science | GPA 3.8 / 4

Champaign, IL

Aug 2024 – May 2026
- Bauman Moscow State Technical University**

Specialist in Autonomous Informational and Control Systems | GPA 4.7 / 5

Moscow, Russia

Sep 2016 – Jun 2022

SKILLS

**Competencies:** ML, deep learning, LLMs, computer vision, statistical modeling, A/B testing, full stack development, HPC  
**Programming Languages:** Python, R, C++, Java, TypeScript, SQL, MATLAB  
**Tools & Frameworks:** PyTorch, LangGraph, vLLM, AWS, Docker, React, Zustand, Tableau, Grafana, PySpark, Neo4j, Weights & Biases