

Mansurbek Satarov

Ph.D. Researcher · Trustworthy AI · LLM Security · Agentic Systems

@sataromk@mail.uc.edu |  linkedin.com/in/mansurbek |  github.com/MansurPro |  Cincinnati, OH

Research Interests

Focus: Trustworthy machine learning; LLM alignment and security; RLHF and learning dynamics; agentic systems and tool security; multimodal learning; scalable ML systems.

Education

University of Cincinnati <i>Ph.D. in Computer Science, fully funded doctoral assistantship</i> - Advisor: Prof. Justin Zhan, Fred A. & Juliet Esselborn Geier Chair Professor and Department Head. - Research focus: LLM alignment, RLHF safety, training-time attacks, learning dynamics, and agent security.	Cincinnati, OH Aug 2025 – Present
North American University <i>B.S. in Computer Science (Software Engineering), GPA: 3.96/4.00</i>	Houston, TX Aug 2022 – May 2025

Research & Industry Experience

Procter & Gamble (P&G), Digital Accelerator at UC <i>Research Assistant, Part-Time</i> - Support cross-functional research through structured data collection, synthesis, quality validation, and stakeholder-ready reporting. - Develop reproducible research and analysis workflows while following data-governance, privacy, and confidentiality requirements.	Cincinnati, OH Jun 2026 – Present
University of Cincinnati <i>Ph.D. Research Assistant, advised by Prof. Justin Zhan</i> - Run controlled fine-tuning experiments across staged training conditions to localize where unsafe behavior forms in preference-based (RLHF) pipelines. - Instrument training pipelines with per-update influence analysis to attribute model behavior to individual training updates and support targeted mitigation. - Develop analysis and policy tooling for governing third-party LLM agent extensions, combining static analysis, permissions, and sandboxed runtime evaluation.	Cincinnati, OH Aug 2025 – Present
University of Cincinnati <i>Teaching Assistant – Python, Machine Learning, and Applied AI</i> Lead labs, grade programming and ML assignments, and help students debug implementations and interpret model behavior.	Cincinnati, OH Spring 2026 – Present
North American University Research Center <i>Research Assistant</i> Built computer-vision and predictive-learning pipelines with PyTorch, CNNs, YOLO, and ViT-Pose on datasets exceeding 10K samples.	Houston, TX Dec 2023 – May 2025
AiKYNETIX LLC <i>Software Engineer / Team Lead</i> Led delivery of ML-enabled Django and React systems; integrated YOLO/ViT-Pose inference and deployed containerized services on NVIDIA GPU servers.	Remote May 2023 – Aug 2025

Selected Projects

- Agent Skill Governance** | *Python, Pydantic, OPA/Rego, Semgrep, OpenAI* 2026 – Present
- Building a research harness for vetting, sandboxing, and monitoring third-party LLM agent skills and MCP servers across their lifecycle; specifics withheld during anonymized review.
- Multi-Agent Deep Research** | *LangGraph, OpenAI, Tavily, Pydantic* | [Repository](#) 2026
- Implemented supervisor/worker orchestration, live search, structured extraction, source-quality scoring, deterministic report synthesis, and resilient no-API fallback paths.
- JARVIS – Rust Open-Source Study** | *Rust, Tauri, Svelte, Vosk* | [Fork](#) 2026 – Present
- Experimenting with an offline voice-assistant codebase to study Rust desktop architecture, speech recognition, wake-word detection, and privacy-preserving local inference.
- Retail Intelligence Platform** | *FastAPI, React, Azure SQL, scikit-learn* | [Repository](#) 2026
- Deployed an end-to-end analytics platform for customer lifetime value, churn risk, basket associations, household search, and interactive reporting, with a persistent SQL-backed dashboard cache.
- GCP College Inquiry Chatbot** | *FastAPI, Gemini, App Engine* | [Repository](#) 2026
- Built a cloud-deployable student inquiry application and a separately scoped Gemini assistant with grounded responses and guardrails for free-text questions.
- Snowflake Financial Analytics** | *Snowpark, Streamlit, Altair* | [Repository](#) 2026
- Built a Snowflake-native dashboard for stock and FX time-series analysis with windowed transformations, interactive filters, and role-based sharing.
- ML & Generative AI:** SmartWaste-Guard (YOLO/ONNX streaming detection), RAG Assistant (ChromaDB retrieval), DocuParse (multilingual PDF-to-Markdown), and Prompt2Clip (Mochi/Diffusers on Modal GPUs).
- Data & Cloud Systems:** Recommender Systems Security (SVD/NMF poisoning), Big Data Weather Analysis (PySpark/NOAA), and Containerized Data Pipeline (Docker/Kubernetes).
- Build Our Own Docker** | *Ongoing systems project* 2026 – Present
- Building a minimal Docker-like container runtime from first principles, focusing on Linux namespaces, cgroups, layered filesystems, image handling, and isolated process execution.
- Additional public work:** [Databricks COVID Impact Analysis](#), [AWS CloudProfile](#), [BeeVision](#), [DigitRecognizer](#), and [DinoMindEvolution](#).

Technical Skills

Languages: Python, JavaScript, TypeScript, Rust, C/C++, Java, SQL, Bash, HTML/CSS

ML/LLM: PyTorch, TensorFlow, Hugging Face Transformers, OpenAI/Gemini APIs, Ollama, RLHF, PPO, DPO, RAG, YOLO, ViT, scikit-learn

AI Engineering Tools: Claude Code, OpenAI Codex, GitHub Copilot, OpenHands, OpenCode, OpenDesign, LangGraph, MCP, reusable agent skills

Agent/Security Research: Pydantic, OPA/Rego, Semgrep, static analysis, sandbox evaluation, tool orchestration, contract enforcement

Systems/Cloud: Docker, Kubernetes, Linux, NVIDIA GPUs, DeepSpeed, Modal, Azure, GCP, Databricks, PySpark, Snowflake/Snowpark

Web/Data: FastAPI, Django, Flask, Streamlit, Node.js, React, Next.js, Tauri, Svelte, PostgreSQL, MySQL, MongoDB, Firebase, ChromaDB

Developer Tools: Git/GitHub, GitHub Actions, Google Colab, Figma, Tailwind CSS, Bootstrap, Material UI, Sanity, Stripe

Leadership & Recognition

President, ACM Club – led hackathons, AI workshops, and collaborative student projects.

Merit-Based Scholarship Recipient (\$100,000), North American University; Fully Funded Ph.D. Scholarship, University of Cincinnati.

Participant: UTSA Hackathon 2023, Rice Datathon 2023, HackRice 2024, and Texas A&M TidalHack 2025.