Mitchell Marfinetz

Phone: (814) 923-7851
Email: mitch@integral.link
LinkedIn: linkedin.com/in/mitchellmarfinetz
GitHub: github.com/0xmarf
Location: Erie, PA

Professional Summary

AI Safety Researcher and ML Engineer with 7+ years developing and evaluating AI systems for safety and alignment. Experienced in Constitutional AI principles, RLHF techniques, and red teaming through autonomous agent development and LLM evaluation projects. Proven track record in empirical AI research, multi-agent coordination, and AI system interpretability. Passionate about building helpful, honest, and harmless AI systems aligned with human values.

Core Technical Skills

AI Safety & Alignment: Constitutional AI, RLHF, AI system evaluation, red teaming, interpretability, safety research
Machine Learning: Large Language Models, reinforcement learning, neural networks, model finetuning, empirical evaluation
Programming Languages: Python, JavaScript, TypeScript, Solidity, SQL
ML Frameworks: PyTorch, TensorFlow, scikit-learn, Pandas, NumPy
Research Methods: Empirical analysis, statistical modeling, A/B testing, experimental design

Professional Experience

AI Safety Researcher & Data Analyst | JEYLabs | Remote | 2023–Present
- Developed autonomous AI agents with safety constraints and monitoring systems, implementing Constitutional AI principles for reliable decision-making in high-stakes environments
- Conducted empirical AI safety research through multi-agent coordination experiments, achieving 20% improvement in agent reliability through systematic evaluation methodologies
- Built red teaming frameworks for DeFi protocols using Python, identifying safety vulnerabilities through automated adversarial testing and threat modeling
- Published technical research on AI-crypto convergence, modeling agent planning and memory systems for safe autonomous operation

Research Engineer & Community Lead | Integral | Remote | March 2022–2023
- Led empirical research on decentralized systems using statistical analysis and ML techniques, conducting systematic evaluation of governance mechanisms affecting 5,000+ participants
- Developed AI safety evaluation frameworks for community moderation, implementing harmlessness principles and bias detection in automated systems

- Performed large-scale data analysis using SQL and Python, segmenting user behavior patterns and identifying safety-critical decision points
- Designed and executed controlled experiments on community engagement, applying empirical research methods to human-AI interaction studies

AI Safety Research Projects

Constitutional AI Agent Framework | BlogWriter Project | 2023–Present
- Implemented Constitutional AI methodology for content generation agents, developing self-critique and revision capabilities aligned with safety principles
- Achieved 40% improvement in output safety and coherence through systematic evaluation and RLHF-inspired feedback loops
- Built comprehensive evaluation suite for AI system behavior, including red teaming protocols and interpretability analysis

Multi-Agent Safety Coordination | Liquidation Bot Research | 2023–Present
- Developed reinforcement learning framework for safe autonomous trading agents with built-in safety constraints and monitoring
- Implemented AI alignment techniques including reward modeling and constitutional training for reliable multi-agent coordination
- Conducted empirical evaluation of agent decision-making under adversarial conditions, contributing to AI safety research methodologies

AI System Interpretability Research | MEV Detection Framework | 2022–Present
- Built machine learning models for AI system behavior analysis and anomaly detection in high-frequency decision environments
- Developed interpretability tools for understanding AI agent decision-making processes, supporting safety evaluation and alignment research
- Applied statistical methods and empirical analysis to validate AI system safety properties under various operational conditions

Research & Publications

AI-Crypto Convergence Research | Technical Blog Series | 2023–Present
- Published research on AI agent coordination in decentralized systems, exploring safety implications of autonomous multi-agent networks
- Contributed to AI safety discourse through empirical analysis of agent behavior in complex environments

Education & Certifications

Certificate in Applied Data Science (Deep Learning Focus) | MIT Professional Education | 2022
- Specialized coursework in neural networks, machine learning safety, and empirical AI research methods

Blockchain Developer Certification | Alchemy University | 2022

Bachelor of Business Administration, Computer Information Systems | Kent State University | 2021
Dean's List Recognition | Fall 2020

Technical Achievements

- Built recommendation systems with 85% accuracy using collaborative filtering and safety-aware design principles
- Developed convolutional neural networks achieving 92% accuracy with robust evaluation methodologies
- Implemented regression models with $R^2 > 0.85$ for predictive analysis in AI safety applications