Julian Quick

Trustworthy Autonomous Systems

julianquick.com • github.com/kilojoules • quectojoules@gmail.com • linkedin.com/in/kilojoules

SELECTED PROJECTS

- **Multi-Turn Jailbreaking & Mechanistic Interpretability**: DPO-trained 3B adversary jailbreaks 8B victim through multi-turn dialogue (26% ASR, JailbreakBench-100). Linear probes detect attacks at AUC 0.97, yet the victim complies. Causal activation steering at layer 16 shifts ASR by ±13pp; the standard refusal direction is orthogonal and inert. Four persistent SAE features point to architectural vulnerabilities.
- **Vibe Code Linting Tool**: Deployed auditing framework to detect unsafe exception handling in LLM-generated Python code. Evaluation reveals that 50-100% of model-generated exception handlers can silently suppress errors. Shipped production-ready CLI tool with PyPI distribution and pre-commit integration. Submitted to COLM 2026.
- **Cross-domain safety transposition**: Adapted ISO 21448 (automotive safety-of-intended-functionality) to LLM safety. SAE-defined safe operating envelope for Llama-3.1-8B achieves AUC 0.955 on adversarial detection and reveals that successful jailbreaks are optimized to resemble benign prompts.

PROFESSIONAL EXPERIENCE

Visiting Researcher **Jan 2026 – Present**
Berkeley Labs, Berkeley, CA

- Shipped adversarial reinforcement learning framework for hardening controllers against sensor attacks; self-play achieves 18× better worst-case robustness without sacrificing clean-environment performance.
- Developed LLM agentic evaluation framework, using the agent to optimize an optimization script based on testing scenarios and evaluated on held-out scenarios. Agent scaffolding explains more variance than model scale: Qwen 32B matches frontier training performance with 2-file context injection (+20 GWh). Frontier model discovers novel non-monotonic optimization schedule no human published.

Researcher **May 2022 – Present**
Risø Wind and Energy Systems, Technical University of Denmark, Roskilde, Denmark

- Designed and shipped a scalable stochastic dispatch optimization algorithm for large autonomous fleets (1,200+ units), reducing mission design time by 95% compared to legacy methods.
- Deployed XGBoost & LSTM-based anomaly detection models to enhance system safety and provide early warnings for mechanical failures, reducing false alarms by 60% in production environments.
- Developed a predictive validation tool that increased scenario coverage by 72%, focusing expensive V&V campaigns on high-risk operational conditions to ensure production safety.
- Led the V&V strategy for a cross-institutional software-in-the-Loop data pipeline, integrating 5 independent simulation platforms to standardize reliability assessments across digital twins.

Systems Engineering Research Assistant **Aug 2017 – Jan 2022**
Turbulence and Energy Systems Laboratory, Boulder, CO

- Designed control strategies that explicitly model sensor uncertainty, increasing system output by 0.5% while reducing the risk of critical mechanical failure by 47%; this work has been deployed in the field and cited over 100 times.

Research Engineer **Jan 2016 – Aug 2017**
National Renewable Energy Laboratory, Golden, CO

- Built a high-performance Python interface to a legacy C++ API with multi-node parallelism to enable V&V of stochastic simulations under uncertainty.

Software Engineering Intern May 2014 – Aug 2014
National Center for Atmospheric Research, Broomfield, Colorado

- Designed and deployed real-time SQL monitoring system with intuitive visual interface, enabling technicians to quickly assess hundreds of live signals; validated the system in operation aboard NCAR research aircraft during field campaign.

EDUCATION

Ph.D., Mechanical Engineering **2022**
University of Colorado, Boulder, CO

Author of 40+ peer-reviewed papers (h-index 14, 650+ citations).