Jonathan R. Reed

AI Safety | Trust & Safety | Safeguards | Responsible AI Deployment

contact@jonathanrreed.com | Dallas, TX | JonathanRReed.com | github.com/JonathanRReed | linkedin.com/in/jonathanrreed0

Sociology student and AI security builder focused on adversarial evaluation, misuse prevention, privacy-conscious deployment, model behavior, and human-centered risk analysis for LLM systems.

Researching and testing LLM behavior since 2019 across adversarial evaluation, agent security, red teaming, and deployable hardening.

Experience

Hello.World Consulting | Founder & AI Safety / Security Engineer | Jul 2022 - Present

- Led AI red teaming, architecture review, and secure deployment work for client LLM systems, focusing on prompt injection, data leakage, tool misuse, and operational failure modes.
- Built private retrieval and model workflows with access controls, logging boundaries, documentation, and human-review steps so teams could adopt AI without losing control of data or process.
- Reduced exploit success and policy bypasses by 35% in tested systems and cut inference costs by 20-45% in selected deployments through local or private-cloud designs.
- Translate technical findings into operational guardrails, launch decisions, and written handoff that make risk ownership clearer after deployment.

Reed & Terry, L.L.P. | IT & Security Lead | Jun 2023 - Present

- Deployed AI retrieval and drafting workflows in a confidentiality-sensitive environment with access controls, least-privilege permissions, patching, and restore-tested backups.
- Managed IAM lifecycle, access reviews, document-handling workflows, validation habits, and secure adoption patterns to protect sensitive legal data without blocking usable AI workflows.
- Helped move secure AI use from concept to real workflow while keeping privacy, human review, and maintainability in scope.

Podium Education | Team Lead, AI Programs | Dec 2025 - Present

- Teach and mentor 400+ students on practical AI use, prompt hygiene, data handling, overtrust, and risk-aware adoption across live sessions and project work.
- Created grading and feedback methods that reduced evaluation turnaround time by 31% while helping the team improve AI system use and rubric consistency.

The Global Career Accelerator | Product Strategy & Prototyping Intern | Aug 2025 - Nov 2025

- Trained cohort members on practical AI use and safety while building demos for L'Oreal, American Eagle, Intel, and charity: water.

Methodist Dallas Medical Center | EMT-B, Emergency Department Intern | Aug 2024 - Dec 2024

- Worked ER and ambulance shifts where risk, documentation, triage, and human impact had immediate consequences.

Selected Projects

PoliBench | Astro, React, Bun, Convex | GitHub: github.com/JonathanRReed/Poli-bench

- Political compass-style benchmark for AI assistants measuring profile, answer stability, parse quality, refusal behavior, cost, latency, war posture, and deviance signals.

RAGFuzz | AI testing | GitHub: github.com/JonathanRReed/RAGFuzz

- Testing project for probing RAG behavior, prompt injection risk, retrieval instability, and failure cases before deployment.

JR AutoRAG | Python, RAG | GitHub: github.com/JonathanRReed/JR-AutoRAG

- Local/private retrieval workflow pattern for document-grounded AI systems with deployable hardening and safer handoff.

AI-Stats | Astro, TypeScript, Supabase | Live: aistats.jonathanrreed.com | GitHub: github.com/JonathanRReed/Ai-stats

- Model comparison dashboard for inspecting provider, cost, context, and benchmark tradeoffs before deployment.

TRACED | Astro | Live: traced.jonathanrreed.com | GitHub: github.com/JonathanRReed/traced

- Security archive and password-exposure tool built around breach data, privacy-preserving checks, and incident context.

Education

The University of Texas at Dallas, B.A. Sociology, Expected 2027 | EMT-B Certificate, UT Dallas/UEMR, Grade A | Wharton County Junior College, Core Curriculum, GPA 3.35

Selected Certifications & Recognition

Multi-time Gray Swan Arena top-five, delivering competition and contracted evaluations, including pre-release, non-NDA testing of GPT-5, Claude Sonnet 4, Grok 4, Llama Maverick, and Kimi K2 Thinking.

IBM AI Engineering Professional Certificate; Vanderbilt Prompt Engineering; Red Teaming for Generative AI; Ethics in the Age of Generative AI; PMI Ethics in Technology; Imperial Mathematics for Machine Learning; Google Cybersecurity coursework.

Skills

Adversarial evaluation; safeguards; prompt-injection testing; misuse prevention; AI governance; trust and safety; RAG evaluation; model behavior analysis; human-in-the-loop rollout; Python; TypeScript; risk analysis; documentation; secure adoption.