Divij Chawla

divijchawla8@gmail.com | LinkedIn | GitHub | +1 (206) 786 8970 | Google Scholar

Focus: AI security ML (automated red teaming, safety benchmarking, monitoring). Built provider-SDK agent harnesses and trace-scored evals to detect vulnerabilities; shipped reproducible PyTorch pipelines with Docker and Slurm/HPC.

Education

University of Washington Seattle, WA

Bachelor of Science in Computer Science *Expected Graduation: 2028*

- GPA: 3.80 | Dean's List | *Pursuing Double Major in Physics*
- **Relevant Coursework:** Mathematical Physics (Phys 227), Linear Algebra, Multivariable Calculus, Programming Languages (CSE 341), Computer Programming III (Java), Project Management Fundamentals (CSE 492F)

Experience

SPAR Research Fellow (Shutdown-Bench) Feb 2026 – Present

Supervised Program for AI Research *Remote*

- Building a safety benchmark for LLM/agent **shutdownability** in realistic tool-use tasks: scenario suite, instruction hierarchy, and failure-mode taxonomy
- Implementing automated red-team agent harness (provider SDK) + trace scoring to detect shutdown resistance (delay/deflect/evasion/goal-preservation); logging tool traces for monitoring

AI Security Researcher March – August 2025

Walled AI *Singapore, Remote*

- **First Author, EMNLP**: Industry Track 2025, Built FINRISKEVAL dataset + eval pipeline (1,720 profiles; 8 models, 13k+ outputs) to measure correctness and intent alignment in finance; ran 100+ prompt/scoring ablations; [Paper]
- Co-authored IMDA-commissioned LLM deployment playbooks: red-teaming guidebooks, domain safety evals, intent alignment and guardrails for production systems; [Link]

AI Research Intern July 2024 – March 2025

University of Bristol (UoB) *Bristol, UK, remote*

- Built eval harnesses for latent alignment failures via representation backdoor attacks; fine-tuned attacker/defender models, tested mitigations, and quantified reliability under white and black box conditions
- Threat-modeled 5+ attack vectors; measured sleeper-agent persistence and mitigation breakpoints across settings

Software Development Intern June 2023 – July 2024

Emsec Private Limited *Bangalore, India, Remote*

- Built and operated a honeypot fleet simulating 7k+ vulnerable applications to capture real-world attacker recon
- Investigated live attacker activity across 10+ honeypots, curating datasets and labeling adversarial activity to strengthen risk-aware defensive automation for 20+ organisations
- Implemented telemetry→evidence pipeline to parse/enrich/aggregate attacker logs into detection signals

DSP (Digital Signal Processing) Intern May – August 2025

Emsec Private Limited *Bangalore, India*

- Optimized C++ DSP modules for high-throughput SDR pipelines; integrated real-time ingestion and processing for low-latency inference, with profiling + latency tuning

Projects

OLMo-core (AllenAI) — Open-Source Pretraining Infrastructure

- Implemented DataMixture MonitorCallback to report per-source token and sequence statistics during pretraining
- Leveraged per-instance metadata to emit aggregated token share & sequences via the trainer logging interface
- Added path-aligned source metadata propagation for mixture datasets; contributed upstream PR to core library

Additional Activities

Lavin Entrepreneurship Program – Selective Venture-Building Uni cohort (MVP, discovery, pitch)
Interactive Intelligence (I2), UW – General Leadership, Leading Curriculum (student-led NeuroAI org)
Husky Satellite Lab (UW) – CDH Team: Developed low-latency C++ systems logic for data optimization

Technical Skills

Python, C++, PyTorch, Transformers; LLM evaluations/benchmarks, data generation/filtering; OpenAI Agents, Google ADK; SQL, Linux/Git/Docker; Slurm/HPC; evaluation harnesses, experiment tracking