Himanshu Mittal

AI Research Engineer | Safety · Evaluation · Agents

himanshu.mittal.dev@gmail.com himanshumittal.com Linkedin Github

Professional Experience

Research Engineer (AI Safety), *Solumn AI* 03/2026 – Present

- Built a benchmark of **3,000+ cyber evaluation environments** (Reinforcement Learning Environments in sandboxed Linux containers, terminal-bench format) for agentic vulnerability localization and automated patch generation. Curated tasks by mining public security advisories and built deterministic verifiers; calibrated difficulty so frontier models reach only pass@8 = 40%, to preserve measurement headroom.
- Led development of **ToolDojo**, a benchmark comprising workflow-driven tasks and 100+ realistic enterprise mock tools that embed prompt injections to evaluate **adversarial tool use** in frontier models.
- Designed an **automated red-teaming framework** that generated compositional jailbreaks using LLM-driven attack recipes combining instruction, obfuscation, and behavioral manipulation strategies.
- Engineered concurrent evaluation infrastructure to efficiently execute frontier safety benchmarks (AgentHarm, HarmBench, CyberSecEval, AirBench, LongFact) under provider API rate limits.

Technical Co-Founder, *Faradworks* 05/2025 – 02/2026

- Architected and deployed a multimodal multi-agent RAG platform on Google Vertex AI for PCB verification over 100+ page engineering documents, **reducing manual verification effort by 90%**.
- Led technical strategy from customer discovery through production deployment, secured pre-seed funding from **Antler**, and ran **50+ enterprise pilots** including deployments with a global semiconductor company.

Data Scientist (LLM Alignment), *Turing* 07/2024 – 04/2025

Generated and quality-controlled **SFT/RLHF** training data for a frontier lab's **pre-release model checkpoints** across reasoning, coding, data analysis, and instruction following. Built response-generation pipelines with conditional prompt chaining validated against gold datasets, graded agent trajectories against policy rubrics, and calibrated LLM-as-a-Judge evaluators to ensure policy adherence.

Founding Data Scientist, *Ripik AI* 10/2022 – 06/2024

Worked directly with **Fortune 500** manufacturers as a forward-deployed AI engineer, deploying Industry 4.0 optimization systems (planning, scheduling, blending) across 15+ manufacturing sites, a **7% production-efficiency gain,** equivalent to adding weeks of annual output at zero capital cost.

Applied ML Engineer 11/2019 – 07/2022

Volvo-Eicher Commercial Vehicles • Bharat Electronics • Times Internet • Datakalp • LeadingIndia AI

Built production ML systems spanning NLP, computer vision, reinforcement learning, and optimization, including job-card analytics, PPE compliance, CTR optimization, and human activity recognition.

Education

B.Tech in Computer Science and Engineering (CGPA = 9.24) 2018 – 2022 | Greater Noida, India

Bennett University (The Times Group of India)

Publications

ML approach to predict net radiation over crop surfaces, *International Journal of Biometeorology* 2022

Skills

Programming (Python, SQL, Linux), **LLMs** (PyTorch, Transformers, SFT, RLHF, LLM Evaluation, VLMs, Frontier Model APIs), **Agentic AI** (RAG, Multi-Agent Systems, Tool Use, MCP, LangGraph), **AI Safety & Evals** (Benchmark Design, Red Teaming, Prompt Injection, Jailbreaks, Adversarial Robustness, Inspect AI, Sandboxed Environments, Synthetic Data Generation), **Infrastructure** (FastAPI, Docker, GCP, AWS, Vertex AI, MLflow)