HARSH RAJ

harsh777111raj@gmail.com
harshraj172.github.io
Google Scholar ◇ Github ◇ LinkedIn

EDUCATION

Northeastern University *Dec 2026 (expected)*
M.S. in Computer Science
GPA: 3.6/4.0

Delhi Technological University (DTU) *August 2023*
B.Tech. in Engineering Physics
GPA: 3.4/4.0

EXPERIENCE

Research Assistant, Forschungszentrum Jülich July 2025 – Present
Remote Cologne, Germany

· Contributing to European Union's Open Foundational Model projects, including MINERVA, OpenEuroLLM, and ELLIOT.

Co-Founder, Ontocord AI Oct 2024 – Present
Remote California, US

· Released the MixtureVitae dataset - the strongest permissibly licensed pretraining dataset to date.

Applied Scientist, Vijil AI Aug 2023 – Oct 2024
Remote California, US

· Developed vijil-fuzzer, an LLM red-teaming framework for enterprise-grade customers. At the launch it got 80% attack rate against gpt-4o.
· Integrated open-source benchmarks into the vijil-evaluation service.

SELECTED PUBLICATIONS

- **Harsh Raj**, Vipul Gupta, Domenic Rosati, Subhabrata Majumdar. *Improving Consistency in Large Language Models through Chain of Guidance* (TMLR'2025)
- Domenic Rosati, Giles Edkins, **Harsh Raj**, David Atanasov, Kai Williams, Subhabrata Majumdar, Janarthanan Rajendran, Frank Rudzicz, Hassan Sajjad. *Defences against Reverse Preference Attacks* (AI Security Workshop, AAAI'25)
- Laura Fee Nern, **Harsh Raj**, Maurice Georgi, Yash Sharma. *On Transfer of Adversarial Robustness from Pretraining to Downstream Tasks* (NeurIPS'23)
- **Harsh Raj**, Domenic Rosati, Subhabrata Majumdar. *Measuring Reliability of Large Language Models through Semantic Consistency* (Best Paper Award, ML Safety Workshop, NeurIPS'22)

PROJECTS AND LIBRARIES

- terminal-bench (600 GitHub ★): Language Agent Sandbox Evaluation
  Co-authored with Mike Meryll, Alex Shaw, Nicholas Carlini
- inspect-ai (1.3k GitHub ★): LLM Evaluation Suite.
  Contributor
- garak (5k GitHub ★): LLM Red-Teaming Suite
  Contributor

AWARDS, GRANTS AND HONORS

SWISS AI Grant: Awarded 200k GPU hours and CHF 61k in funding. *2025*
Vision and Language Navigation: Secured 3rd position on the R2R benchmark *2024*
ML Safety Challenge: Best Paper Award with a $5000 prize *2022*
Codeforces ML Round: Rank 7 in the Raif ML Round 1 *2021*