Aman Priyanshu

Post-Training Expert | Foundation Models for Security

amanpriyanshu.github.io @ amanpriyanshusms2001@gmail.com github.com/AmanPriyanshu
Google Scholar twitter.com/AmanPriyanshu6 linkedin.com/in/Aman-Priyanshu

Professional Experience

Present – Jan 2025 | **Foundation-AI (Cisco Systems, Inc.) | AI Researcher** | **In-Person / San Francisco, CA, USA**
Developed LLMs for security focused on reasoning, recursive tool usage, and agentic behavior. Specialized in post-training, test-time scaling, & bandit-style diverse path exploration for automated pen-testing.

Aug 2024 – Jun 2024 | **Robust Intelligence | AI Security Research Intern** | **In-Person / San Francisco, CA, USA**
Jailbroke LLaMA-3.1(499×) & OpenAI(4.25× Attack Success Rate) in 24h [received media coverage]; Developed automated prompt-injections; Created million-scale harmful intent dataset for AI Safety.

Mar 2024 – Jan 2024 | **OpenAI | Red Teaming Network, Independent Contractor** | **Remote / San Francisco, CA, USA**
Participated in OpenAI led red teaming efforts to assess the risks and safety profile of OpenAI models. (NDA)

Oct 2024 – Dec 2025 | **Anthropic | Bug Bounty Program, Independent Contractor** | **Remote / San Francisco, CA, USA**
Participated in Anthropic's Bug Bounty Program (NDA). Discontinued following graduation.

Education

Dec 2024 – Aug 2023 | **Carnegie Mellon University** | **Pittsburgh, PA, USA**
MSIT — Privacy Engineering

May 2023 – Jul 2019 | **Manipal Institute Of Technology, MAHE** | **Karnataka, India**
B.Tech Information Technology *with Minors in Big Data Analytics*

Research Experience

May 2024 – Aug 2023 | **Privacy Engineering Research** [🌐] | **Carnegie Mellon University, USA**
Independent Study | Advisor: Professor Norman Sadeh
Project: For prompt-engineering geared towards usable privacy & security.

Aug 2023 – Mar 2023 | **OpenMined | Research Team** [🌐] | **Remote / United Kingdom**
Project Lead and Collaborator | Collaborators: Dr. Niloofar Mireshghallah
Project: The impact of epsilon differential privacy on LLM hallucinations.

Honours and Awards

- Spark Grant Winner, NOVA Hackathon, Mar 2024
- Theme Category Winner, HackCMU, Sept 2023
- Second Runners-Up - ShowYourSkill (Coursera), Jun 2022
- AAAI Undergraduate Consortium Scholar, Feb 2023
- First Prize - HackRx by Bajaj Finserv, July 2021
- First Prize - ACM UCM Datathon, UC Merced, May 2021

Publications

S=In Submission, J=Journal, C=Conference, (* = Equal Contribution)

[TR.1] **Llama-3.1-FoundationAI-SecurityLLM-Reasoning-8B Technical Report** [Preprint]
arXiv Preprint [**arXiv'26**]

[C.2] **What Lies Beneath the Guardrails? Jailbreaking Meeting Bias Audit**
2025 AAAI Conference on Artificial Intelligence [**AAAI'25**]

[C.1] **When Neutral Summaries are not that Neutral: Quantifying Political Neutrality in LLM-Generated Summaries**
2025 AAAI Conference on Artificial Intelligence [**AAAI'25**]

[S.1] **Are Chatbots Ready for Privacy-Sensitive Applications? An Investigation into Input Regurgitation and Prompt-Induced Sanitization** [Preprint]
[In Submission]

Other venues of acceptances: AI4SG@AAAI'23, UpML@ICML'22, IEEE S&P'21, RCV@CVPR'21, and W-NUT@EMNLP'21.

Skills

Programming Languages Python, Java, Go, C++, C, C#, SQL, Shell Scripting (Git & Bash)
Frameworks & Libraries PyTorch, Tensorflow, JAX, HuggingFace, FastAPI, AdaptKeyBERT & NERDA-Con (self-authored)
Relevant Coursework Prompt Engineering (17730), AI Governance (17716), Deep Learning (11785), Computer Technology Law (17562), Differential Privacy (17731), Information Security (17631), & Usability (17734).