

# Aman Priyanshu

## Post-Training Expert | Foundation Models for Security

🌐 [amanpriyanshu.github.io](https://amanpriyanshu.github.io) 📲 [amanpriyanshusms2001@gmail.com](mailto:amanpriyanshusms2001@gmail.com) 🐧 [github.com/AmanPriyanshu](https://github.com/AmanPriyanshu)  
🎓 [Google Scholar](#) 🐦 [twitter.com/AmanPriyanshu6](https://twitter.com/AmanPriyanshu6) 💬 [linkedin.com/in/Aman-Priyanshu](https://linkedin.com/in/Aman-Priyanshu)

## Professional Experience

Present	<b>Foundation-AI (Cisco Systems, Inc.)</b>   AI Researcher	In-Person / San Francisco, CA, USA
Jan 2025	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	<b>Robust Intelligence</b>   AI Security Research Intern	In-Person / San Francisco, CA, USA
Jun 2024	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	<b>OpenAI</b>   Red Teaming Network, Independent Contractor	Remote / San Francisco, CA, USA
Jan 2024	Participated in OpenAI led red teaming efforts to assess the risks and safety profile of OpenAI models. (NDA)	
Oct 2024	<b>Anthropic</b>   Bug Bounty Program, Independent Contractor	Remote / San Francisco, CA, USA
Dec 2025	Participated in Anthropic's Bug Bounty Program (NDA). Discontinued following graduation.	

## Education

Dec 2024	<b>Carnegie Mellon University</b>	Pittsburgh, PA, USA
Aug 2023	MSIT — Privacy Engineering	
May 2023	<b>Manipal Institute Of Technology, MAHE</b>	Karnataka, India
Jul 2019	B.Tech Information Technology with Minors in Big Data Analytics	

## Research Experience

May 2024	<b>Privacy Engineering Research</b> [🔗]	Carnegie Mellon University, USA
Aug 2023	<i>Independent Study / Advisor: Professor Norman Sadeh</i> <u>Project:</u> For prompt-engineering geared towards usable privacy & security.	
Aug 2023	<b>OpenMined   Research Team</b> [🔗]	Remote / United Kingdom
Mar 2023	<i>Project Lead and Collaborator / Collaborators: Dr. Niloofar Mireshghallah</i> <u>Project:</u> The impact of epsilon differential privacy on LLM hallucinations.	

## Honours and Awards

- Spark Grant Winner, NOVA Hackathon, Mar 2023
- 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-8B Technical Report** [Preprint]  
arXiv Preprint [arXiv'25]
- [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

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