Aman Priyanshu

Graduate Student at Carnegie Mellon University

🌐 amanpriyanshu.github.io @ apriyans@andrew.cmu.edu github.com/AmanPriyanshu
🎓 Google Scholar twitter.com/AmanPriyanshu6

Education

Present
Aug 2023 | **Carnegie Mellon University** — **Pittsburgh, PA, USA**
MSIT — Privacy Engineering

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

Professional Experience

Present
Jun 2024 | **Robust Intelligence | AI Security Research Intern** — **In-Person / San Francisco, CA, USA**
Jailbroke LLaMA-3.1 & OpenAI within 24 hours each, exploits disclosed & gained media coverage; Developed automated prompt-injections; Created million-scale harmful intent dataset for AI safety compliance.

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.

Aug 2023
Aug 2022 | **Eder Labs R&D Private Limited | Privacy Engineer Intern** — **Hybrid / Delaware, USA**
Worked on differentially private synthetic data generation & privacy-preserving recommendation systems.

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.

Present
Aug 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.

Aug 2022
Jun 2022 | **Concordia University** [🌐] — **Montreal, Canada**
MITACS Globalink Research Intern | Advisors: Professor Wahab Hamou-Lhadj
Project: Exploring machine learning for anomaly detection toolkit.

Honours and Awards

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

Publications

S=In Submission, J=Journal, W=Workshop, (* = Equal Contribution)

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

[C.1] **Through the Lens of LLMs: Unveiling Differential Privacy Challenges** [Presentation]
2024 USENIX Conference on Privacy Engineering Practice and Respect **[PEPR@USENIX'24]**

[J.1] **Finding an elite feature for (D)DoS fast detection-Mixed methods research** [PDF]
Journal: Computers & Electrical Engineering, Volume: 98, Pages: 107705, 2021 **[Computers & Electrical Engineering'21]**

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, SQL, Shell Scripting(Git & Bash)
Frameworks	PyTorch, Tensorflow, JAX, NLTK, Huggingface, FastAPI, Flask
Relavant Coursework	Prompt Engineering (17730), AI Governance (17716), Deep Learning (11785), Computer Technology Law (17562), Differential Privacy (17731), Information Security (17631), Usability (17734), Data Structures & Algorithms, OOPs, and Database Management.