Yash Maurya

maurya.yash152@gmail.com | yashmaurya.com | linkedin.com/in/yashmaurya | Google Scholar | San Francisco, CA

EXPERIENCE

Scale AI — Pittsburgh, PA
Research Engineer — May 2025 - Present

- Led **LLM red teaming** and **adversarial testing** projects targeting **safety vulnerabilities** for frontier labs like Google, Deepmind, xAI, OpenAI
- Built multiagent data evaluation pipeline to grade human responses, improving turnaround time from 1-2 weeks to **~12 min** with **90+%** precision
- Developing automated red-teaming framework to evaluate enterprise LLMs against privacy policies and content guidelines

Bank of New York Mellon (BNY) — Pittsburgh, PA
AI Governance Intern — June 2024 - Aug 2024

- Architected **model evaluation pipeline** for LLMs (Mixtral, Llama-2, GPTs) benchmarking across **accuracy, safety, and fairness** metrics
- Built real-time PII detection system integrating multiple pre-trained models via **Microsoft Presidio** and **NVIDIA NeMo** frameworks
- Conducted NLP analytics on platform usage **(15,000+ users)** using clustering & topic modeling to develop **risk-based evaluation** strategies
- Developed **LLM guardrails** and automated testing framework using industry benchmarks (MMLU, GSM8k, SALAD-Bench, RAGAS, etc)

Samsung Electronics — Noida, India
Research Engineer — July 2022 - Aug 2023

- Developed image narrative generation system using EfficientPS and UPSNet for panoptic segmentation, enhancing Samsung Discover 2.0
- Engineered large-scale data pipeline using Selenium and Beautiful Soup, processing and cleaning **100k+ news articles daily**
- Built unsupervised topic taxonomy system for **10M+ articles** powering Samsung News recommendations, optimizing content discovery

DynamoFL (YC W22) — San Francisco, CA | Remote
Federated Learning Researcher — Feb 2021 - Aug 2021

- Developed production-grade federated learning algorithms (**FedAvg, FedProx, FedMD, FedHE**) with focus on distributed model training
- Implemented **differential privacy** mechanisms using PyDP, evaluating noise injection methods for private model training
- Built synthetic data generation system combining PII detection (Microsoft Presidio) and tabular synthesis (CTGAN) for ML training

EDUCATION

Carnegie Mellon University (CMU) — Pittsburgh, PA
Master of Science in Information Technology - **Privacy Engineering** (MSIT-PE) | CGPA 3.9 / 4.0 — Dec 2024
Awards: **IAPP Westin Scholar 2024**, CMU Spark Entrepreneurship Grant Winner
Graduate Courses: *Federated Learning, Differential Privacy, AI Governance, Responsible AI, Usable Privacy & Security*
Research Areas: Unlearning in LLMs, Fairness, PETs(Privacy Enhancing Technologies), Synthetic Data, Implicit Bias Auditing

CERTIFICATIONS

Certified Information Privacy Technologist (CIPT), International Association of Privacy Professionals (IAPP) [Credential] — Jan 2024
Privacy Management Professional, OneTrust [Credential] — Feb 2025
AI Security & Governance, Securiti [Credential] — Feb 2025

PROJECTS

UsersFirst: A User-Centric Privacy Threat Modeling Framework for Notice and Choice | Collaboration with **PwC** [Link] — Jan 2024 - May 2024

- Pioneered a novel threat modeling framework that addresses **AI privacy vulnerabilities** in **user interface design**
- Conducted in-depth **interviews** with **20 participants**, validating framework efficacy vs. LINDDUN and PANOPTIC
- Integrated Privacy-by-Design principles to create actionable guidelines for combating deceptive design practices
- Accepted at Symposium of Usable Privacy and Security (**SOUPS 2024**).

Unmasking Threats in Google's Topics API (Replacement of Ad Cookies) | Presented at **USENIX PEPR'24** [Link] — Sept 2023 - Dec 2023

- Calculated Topics API's epsilon(privacy leakage budget) at 10.4 per week (epsilon > 10 signifies inadequate privacy protection)
- Our LLM based on **Hierarchical BERT** achieved **95.41% accuracy** and **86.73% specificity f**or Membership Inference Attacks(MIA)
- Achieved **68.19% re-identification** on an anonymized German Browsing Dataset, far surpassing Google's 1% claim

SELECTED PUBLICATIONS

- **Making Privacy-Preserving AI Accessible: A Practitioner-Oriented Framework, USENIX** Symposium of Usable Privacy and Security **2025**
- **When Privacy Guarantees Meet Pre-Trained LLMs: A Case Study in Synthetic Data, USENIX** PEPR'25 [Link]
- **Position: LLM Unlearning Benchmarks are Weak Measures of Progress**, Secure and Trustworthy Machine Learning **(SaTML) 2025** [PDF]
- **Guardrail Baselines for Unlearning in LLMs**, Secure and Trustworthy Large Language Models Workshop at **ICLR 2024** [PDF]
- **Federated Learning for Colorectal Cancer Prediction**, **2022 IEEE** 3rd Global Conference for Advancement in Technology (GCAT) [PDF]
- **Improved variants of Score-CAM via Smoothing and Integrating**. Responsible Computer Vision Workshop at **CVPR 2021** [Poster]

SKILLS

Programming Languages: Python, Java, C/C++, JavaScript, SQL, Rust, Bash
Libraries/Frameworks : PyTorch, TensorFlow, HuggingFace, OpenAI, Scikit-learn, Numpy, PySyft, Flower, Opacus, OpenDP, Nvidia NeMO
MLOps Tools & Frameworks: Wandb, Mlflow, Optuna, ZenML, Flask, Django, GCP, AWS, Docker, Langchain, W&B, Node.js, Neo4j, Airflow
Privacy Frameworks & Standards: NIST Privacy Framework, LINDDUN, MITRE PANOPTIC, FIPPs, OWASP, Privacy-by-Design, NIST AI RMF