

Yash Maurya

maurya.yash152@gmail.com | yashmaurya.com | linkedin.com/in/yashmaurya | [Google Scholar](https://scholar.google.com/citations?user=...) | Pittsburgh, PA

EDUCATION

Carnegie Mellon University (CMU)

Master of Science in Information Technology - Privacy Engineering (MSIT-PE) | CGPA 3.9 / 4.0

Graduate Courses: *Federated Learning, Differential Privacy, Prompt Engineering, AI Governance, Responsible AI*

Research Areas: Unlearning in LLMs, Fairness, PETs(Privacy Enhancing Technologies), Synthetic Data, Implicit Bias Auditing

Pittsburgh, PA

Dec 2024

EXPERIENCE

Meta

Privacy Researcher

Pittsburgh, PA

Aug 2024 - Dec 2024

- Led data-driven research analyzing user engagement patterns across **500 participants** through experimental interface designs
- Engineered and deployed **full-stack React application** with instrumented logging to evaluate feature adoption and control preferences
- Conducted **statistical analysis** using Python to identify significant correlations between interface design and user interaction patterns
- Built automated data pipeline for processing **behavioral metrics** and survey responses, deriving actionable product insights

Bank of New York Mellon (BNY)

AI Governance Intern

Pittsburgh, PA

June 2024 - Aug 2024

- Architected **Langchain 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

R&D Engineer

Noida, India

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 BeautifulSoup, 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)

Federated Learning Researcher

San Francisco, CA | Remote

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 privacy-preserved model training
- Built synthetic data generation system combining PII detection (Microsoft Presidio) and tabular synthesis (CTGAN) for ML training

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

Guardrail Baselines for Unlearning in Large Language Models

Jan 2024 - May 2024

- Demonstrated that zero-shot prompting can achieve competitive unlearning performance on unlearning benchmarks without fine-tuning
- Extending the baseline by 16-bit/8-bit **quantized fine-tuning** LLaMA-2-7B using **LoRA** and **QLoRA** techniques for efficient unlearning

UsersFirst: A User-Centric Privacy Threat Modeling Framework for Notice and Choice | Collaboration with PwC

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 Usable Privacy and Security (**SOUPS 2024**).

Unmasking Threats in Google's Topics API (Replacement of Ad Cookies) | Presented at USENIX PEPR'24

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** for Membership Inference Attacks(MIA)
- Achieved **68.19% re-identification** on an anonymized German Browsing Dataset, far surpassing Google's 1% claim

SELECTED PUBLICATIONS

- **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