Yash Maurya

maurya.yash152@gmail.com | yashmaurya.com | linkedin.com/in/yashmaurya | Google Scholar | Pittsburgh, PA

EDUCATION

Carnegie Mellon University (CMU) Pittsburgh, PA
Master of Science in Information Technology - Privacy Engineering (MSIT-PE) | CGPA 3.9 / 4.0 Dec 2024
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

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
Privacy Assessments & Documentation: Data Protection Impact Assessments(DPIAs), ROPAs, PIAs, Consent Management, Data Flow Mapping

WORK EXPERIENCE

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

- Implemented LLM guardrails using Microsoft Presidio and NVIDIA NeMo for PII de-identification and content moderation
- Developed anti-jailbreaking rails using local LLMs like Llama-guard and ShieldGemma, enhancing AI system security and compliance
- Engineered automated pipelines to evaluate LLMs and RAG agents using benchmarks like MMLU, GSM8k, SALAD-Bench, and RAGAS
- Contributed to governance validation of Eliza, BNY's AI platform (15,000+ users), collaborating - Risk, Legal, Privacy, & Engineering teams

Samsung Electronics Noida, India
R&D Engineer July 2022 - Aug 2023

- Developed image narrative generation module using EfficientPS and UPSNet for panoptic segmentation in Samsung Discover 2.0
- Built large-scale data extraction, processing & ingestion engine for news articles using Selenium, BS4, processing 100k+ articles daily
- Engineered Unsupervised Topic Taxonomy construction pipeline for 10+ Million articles for Samsung News' recommendation system

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

- Implemented state-of-the-art Federated Learning(FL) algorithms from scratch including FedAvg, FedProx, FedMD, and FedHE
- Evaluated DP techniques like Laplacian and Gaussian noise algorithms using PyDP and prior-independent auctions for federated learning
- Engineered a PII sanitization portal leveraging Microsoft Presidio API and CTGAN for generating clean synthetic tabular data

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
- Accepted at Secure and Trustworthy LLM(SetLLM) Workshop at **ICLR 2024**

UsersFirst: A User-Centric Privacy Threat Modeling Framework for Notice and Choice 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**).

Is it worth storing historical gradients to identify targeted attacks in Federated Learning? | CMU Sept 2023 - Dec 2023

- Improved label flip attack detection by up to 25% in FedAvg using current weights, not historical gradients for N=20,50,100 clients.
- Achieved an improvement of up to 15% for targeted attack detection in FedAvg with Differentially Private-SGD(DP-SGD) integration.
- Promotes data minimization for improving privacy of users and overall reducing storage costs.

Unmasking Threats in 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

CERTIFICATIONS

Certified Information Privacy Technologist (CIPT) | IAPP - International Association of Privacy Professionals | Credential Jan 2024

SELECTED PUBLICATIONS

Wang, T., Li, X. A., Rivera-Lanas, M., **Maurya, Y.**, Habib, H., Cranor, L. F., & Sadeh, N. "UsersFirst: A User-Centric Privacy Threat Modeling Framework for Notice and Choice" **SOUPS 2024**. https://www.usenix.org/system/files/soups2024_poster49_abstract-wang_final.pdf

P. Thaker, **Y. Maurya**, and V. Smith, "Guardrail Baselines for Unlearning in LLMs," **SET LLM@ICLR 2024**. https://arxiv.org/abs/2403.03329

Y. Maurya, P. Chandrahasan and P. G, "Federated Learning for Colorectal Cancer Prediction," 2022 **IEEE** 3rd Global Conference for Advancement in Technology (GCAT), pp. 1-5, doi: 10.1109/GCAT55367.2022.9972224

Rakshit Naidu, Soumya Kundu, Shamanth R Nayak K, **Yash Maurya**, Ankita Ghosh. "Improved variants of Score-CAM via Smoothing and Integrating". **Responsible Computer Vision(RCV) Workshop at CVPR 2021**. 10.13140/RG.2.2.23611.54563.