Yash Maurya

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

EDUCATION

Carnegie Mellon University (CMU) Pittsburgh, PA
Master of Science in Information Technology - Privacy Engineering (MSIT-PE) | CGPA 3.95 / 4.0 Dec 2024
Graduate Courses: *Federated Learning, Differential Privacy, Prompt Engineering, AI Governance*
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, Streamlit, Node.js

WORK EXPERIENCE

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

- Built LLM guardrails(on-prem and cloud) for PII De-identification, Content Moderation, Off-topic rails, & Anti-jailbreaking rails
- Engineered automated benchmarking pipelines for foundation LLMs and RAG agent evaluation for performance, safety, ethics, and risk
- Contributed to governance validation of Eliza, BNY's AI platform (15,000+ users), collaborating - Risk, Legal, Privacy, & Engineering teams
- Our validation of Eliza resulted in a feature in Fortune magazine, marking the platform's first public announcement.

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

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

Samsung Electronics Noida, India
R&D Intern Feb 2022 - June 2022

- Developed an efficient LSTM-based network for next-activity prediction, optimized for on-device mobile deployment
- Designed a ResNet-based CNN to predict COVID-19 from cough sounds by analyzing MFCC images, achieving 83% accuracy

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

- Implemented multiple state-of-the-art Federated Learning algorithms from scratch including FedAvg, FedProx, FedMD, and FedHE
- Evaluated epsilon values for various differential privacy techniques with novel Laplacian and Gaussian noise addition algorithms
- 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 prompting can achieve competitive unlearning performance on popular 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**

Unified Locational Differential Privacy Framework Jan 2024 - April 2024

- Developed a DP framework for private aggregation of diverse geographical data, inspired by Apple's "Learning Iconic Scenes with DP"
- Evaluated on simulated datasets for contagion tracking, income statistics, and voting, ensuring real-world applicability
- Leveraged Opacus library for privacy budget tracking and composition, guaranteeing strong differential privacy

Prompt-Driven Synthetic Data Augmentation for Bias Correction with Differential Privacy Alternative Jan 2024 - March 2024

- Developed a secure data interface leveraging Streamlit, enabling efficient bias detection in datasets with Python, regex, and Sentence-BERT
- Utilized LLMs to generate and apply regex queries for precise bias detection, enhancing fairness in machine learning models
- Created synthetic counterfactuals using GPT-3.5, balancing datasets while preserving data privacy with differential privacy techniques

Unmasking Threats in Topics API (Replacement of Ad Cookies) | CMU 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
- Accepted at **USENIX PEPR'24**

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.