Praney Goyal

✉ praneyygoyal@gmail.com | 🌐 praneyg.github.io | github.com/praneyg

Education

The Pennsylvania State University *May 2025*

Master & Bachelor (Honors) of Science in Computer Science, Minor in Mathematics

Experience

DL/AI Researcher Jul 2025 – Present

Brown University *Remote*

- Investigated vulnerabilities in parameter-efficient fine-tuning, showing how single spurious tokens can manipulate predictions and expose efficiency–robustness tradeoffs
- Conducted systematic experiments across model architectures and datasets to study spurious correlations under varied training and data corruption conditions
- Evaluated mitigation approaches such as grammar checkers and preprocessing, identifying fundamental limitations in existing defenses against adversarial patterns
- Developed diagnostic frameworks and reproducible tools for robustness testing, contributing to AI safety research on secure deployment

Machine Learning Research Assistant Dec 2023 – Mar 2025

RAISE Lab, Pennsylvania State University *University Park, PA*

- Engineered a GPT-4o counterfactual explanation framework that enhanced AI transparency by enabling non-technical stakeholders to comprehend NLP model decisions, leading to a 35% increase in trust scores.
- Boosted model accuracy by 10.4% by integrating counterfactual data augmentation across six machine learning models and three datasets, demonstrating the effectiveness of this approach in improving predictions
- Established new industry benchmarks for explainability by creating an evaluation methodology combining automated metrics and LLM-based validation, setting a standard for assessing model interpretability

Software Developer Aug 2022 – Dec 2022

Exacta Global Smart Solutions *Philadelphia, PA*

- Led a 4-member team to develop a cloud-connected IoT traffic management system using ESP32 and AWS IoT Core, improving efficiency by 40% and earning 3rd place among 50+ teams in the KETI oneM2M Hackathon
- Automated AWS infrastructure with Terraform and CloudFormation, reducing setup time by 60% and ensuring 99.5% uptime through a fault-tolerant design

Machine Learning Engineer Jun 2021 – Aug 2021

Snapdeal *Remote*

- Built a forecasting model using Python and Scikit-Learn to predict server I/O operations, improving prediction accuracy by 40% and supporting efficient resource allocation
- Integrated predictive models with Linux-based server operations, reducing response times by 20% and optimizing system performance

Graduate Teaching Assistant Aug 2022 – May 2024

Pennsylvania State University *University Park, PA*

- Led discussion sections and tutored students about Discrete Mathematics concepts like logic, set theory, and graph theory, improving comprehension and engagement
- Implemented lesson plans and managed classroom dynamics for 1000 students, using educational technology to enhance learning

Publications and Preprints

2025: MM Salles*, P Goyal* *et al. LORA users beware: A few spurious tokens can manipulate your finetuned model.* 2025. **arXiv:** 2506.11402 [cs.LG]. **URL:** https://arxiv.org/abs/2506.11402

2025: MM Salles*, P Goyal* *et al. Paraphrasing Away Malicious Tokens: Improving LLM Finetuning Safety by Filtering Spurious Correlation.* NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle. Poster presentation. **URL:** https://openreview.net/forum?id=7fM4Q0TLgZ

Technical Skills

Languages: Java, Python, C/C++, SQL, JavaScript, HTML/CSS
Technologies: MongoDB, Express, React, Node.js, Flask, Git, Linux, Docker, Amazon Web Services
Libraries: Pandas, NumPy, TensorFlow, Scikit-learn, Matplotlib