JAMES BEETHAM
(515) 520 8082 | jamesjbeetham@gmail.com

Objective: Gain experience with industry research over the course of an internship and evaluate fit for working full time after graduation.

I. Education

University of Central Florida, Orlando, FL — 2020-Present

Pursuing a PhD in Computer Science in the Center for Research in Computer Vision (CRCV)
Expected Graduation: May 2026 — Advisor: Dr. Mubarak Shah

Rutgers University, New Brunswick, NJ — 2017-2020

BS in Computer Science, BA in Cognitive Science, minors in Math and Psychology

II. Publications

Beetham, James, et al. LIAR: Leveraging Alignment (Best-of-N) to Jailbreak LLMs in Seconds. arXiv preprint arXiv:2412.05232 (2024).

Beetham, J., Kardan, N., Mian, A., & Shah, M. Dual Student Networks for Data-Free Model Stealing. *The Eleventh International Conference on Learning Representations* (ICLR 2023).

Beetham, J., Kardan, N., Mian, A., & Shah, M. Detecting Compromised Architecture/Weights of a Deep Model. *In 2022 26th International Conference on Pattern Recognition* (ICPR 2022). IEEE.

III. Related Experience

Jailbreaking LLMs — Spring 2024-Present

Designed an efficient method for automatically bypassing safety protocols in Large Language Models to help improve safety alignment. Work is currently under review for publication.

Video Geo-Localization Research Using GIS — Fall 2022-Summer 2024

Developed an analytical approach which leverages a modified DETR object detector and a graph search to perform geo-localization on videos.

Attack Detection Research for DARPA — Fall 2020-Spring 2022

Designed new methods for detecting adversarial attacks and their tool chains. This work led to a paper published in ICPR 2022 on detecting white-box from black-box attacks.

Mentor in Research Experience for Undergraduates at UCF (REU) — Summer 2021, 2022, 2024

Performed research with undergraduates on various projects. (2021) Explored salient attacks by traversing the latent-space of various GAN architectures. (2022) Detect and defend against data-free model stealing attacks. (2023) Bypass data-free model stealing defenses.

Software Engineering Internship at Okta — Summer 2019

Worked with the Sync Team to implement a function which syncs users missed by the initial provisioning job. Worked in the monolith codebase to write new Java code.

IV. Skills

Languages: Python, Java, JavaScript (Node, TypeScript)
Tools: PyTorch, distributed training (SLURM), Docker, Git version control