Blake Bullwinkel

blakebullwinkel@gmail.com blakebullwinkel.com GitHub LinkedIn Scholar

Education

Harvard University Cambridge, MA
M.S. in Data Science. GPA 3.95/4 *May 2022*

Williams College Williamstown, MA
B.A. in Mathematics, Chinese. GPA 3.83/4 (*cum laude*) *June 2020*

University of Oxford Oxford, UK
Attended as part of the selective, year-long Williams-Exeter Program at Oxford (WEPO). *June 2019*

Professional Experience

Microsoft Redmond, WA
AI Security Researcher II, AI Red Team *Jan 2024–Present*

- Leading research into a variety of GenAI safety and security topics (jailbreaks, prompt injection attacks, model backdoors, etc.) to inform Microsoft's understanding of the AI risk landscape.
- Conducting red team operations to identify vulnerabilities in high-profile Microsoft and OpenAI products (GPT-5, Deep Research, Phi series, etc.) and inform safety mitigations.
- Contributing to PyRIT, an open-source Python framework for identifying risks in GenAI systems.

Data & Applied Scientist *Aug 2022–Dec 2023*

- Introduced a method to classify performance bugs and customer incidents using text embeddings.
- Built a pipeline to detect and prioritize kernel-mode memory leaks across the Azure fleet.

Harvard University Cambridge, MA
Teaching Fellow *Feb–May 2022*

- Assisted professors in teaching of CS 109b: Advanced Topics in Data Science, a course focused on non-linear statistical methods and deep learning models, including CNNs, RNNs, LSTMs, GANs, and transformers.

Recent Research

B Bullwinkel et al. A Representation Engineering Perspective on the Effectiveness of Multi-Turn Jailbreaks. *ICML Workshop on Data in Generative Models, 2025.*

B Bullwinkel et al. Steering Language Model Refusal with Sparse Autoencoders. *ICML Workshop on Actionable Interpretability, 2025.*

B Bullwinkel et al. A Systemization of Security Vulnerabilities in Computer Use Agents. *ICML Workshop on Computer Use Agents, 2025.*

B Bullwinkel et al. Lessons From Red Teaming 100 Generative AI Products. *Microsoft BlueHat 2024. NeurIPS Workshop on Red Teaming GenAI, 2024.*

B Bullwinkel et al. Phi-3 Safety Post-Training: Aligning Language Models with a "Break-Fix" Cycle. *Arxiv 2024.*

B Bullwinkel et al. PyRIT: A Framework for Security Risk Identification and Red Teaming in Generative AI Systems. *CAMLIS 2024.*

Honors and Awards

CES Infinite Mindset Partnership Award for leading Phi-3 language model red teaming (Microsoft) *2024*
Quality Stars Award for building a novel memory leak detection pipeline for Azure (Microsoft) *2023*
Certificate of Distinction in Teaching based on student ratings (Harvard University) *2022*
IACS Student Scholarship to support data science thesis research (Harvard University) *2021*
Goldberg Prize in Mathematics for the best senior mathematics colloquium (Williams College) *2020*
Linen Prize in Chinese for achieving distinction in Chinese (Williams College) *2020*

Skills

Programming	Python, R, HTML/CSS, JavaScript, SQL, KQL
Libraries	NumPy, Pandas, SciPy, Scikit-Learn, PyRIT, HuggingFace, PyTorch, TensorFlow
Platforms	Azure, AWS, Docker, Linux, Windows
Language	Working proficiency in written and spoken Chinese (Mandarin)