Blake Bullwinkel

Contact Information

✉ blakebullwinkel@gmail.com in linkedin.com/in/blakebullwinkel
🌐 blakebullwinkel.com github.com/blakebullwinkel

Education

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

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

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

Publications

Link to Google Scholar profile.

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.* Arxiv 2024.

Z Ma*, S Su*, N Zhao*, L Bieske, **B Bullwinkel**, J Gao, G Liao, S Li, Z Luo, B Wang, Z Wen, Y Yang, Y Zhang, C Bruderlein, W Pan. *Using Large Language Models for Humanitarian Frontline Negotiation: Opportunities and Considerations.* ICML Workshop on the Next Generation of AI Safety (NextGenAISafety), 2024.

R Pellegrin*, **B Bullwinkel***, M Mattheakis, P Protopapas. *Transfer Learning with Physics-Informed Neural Networks for Efficient Simulation of Branched Flows.* NeurIPS Workshop on Machine Learning and the Physical Sciences, 2022.

B Bullwinkel*, D Randle*, P Protopapas, D Sondak. *DEQGAN: Learning the Loss Function for PINNs with Generative Adversarial Networks.* ICML Workshop on AI for Science (AI4Science), 2022.

B Bullwinkel, K Grabarz, L Ke, Sc Gong, C Tanner, J Allen. *Evaluating the Fairness Impact of Differentially Private Synthetic Data.* ICML Workshop on Theory and Practice of Differential Privacy (TPDP), 2022.

Research Experience

AI and Humanitarian Negotiation, Harvard University Sept 2023–June 2024
Capstone Research Course. Advisors: Weiwei Pan, Claude Bruderlein

- Advised an interdisciplinary team of researchers to develop and evaluate LLM-based tools for frontline humanitarian negotiators.

Multimodal Adversarial Attacks, Harvard University Sept 2023–Dec 2023
Capstone Research Course. Advisors: Siddarth Swaroop, Weiwei Pan, Finale Doshi-Velez

- Advised research focused on understanding gradient-based adversarial attacks against Vision Language Models (VLMs).

Physics-Informed Neural Networks, Harvard University Feb 2021–May 2022
Master's Thesis. Advisors: Pavlos Protopapas, David Sondak

- Developed a GAN-based method for obtaining accurate solutions to a wide range of ordinary and partial differential equations.
- Implemented multi-head architectures and transfer learning algorithms to more efficiently simulate branched flows, a universal wave phenomenon.
- Maintained research code in a user-friendly PyTorch package.

Interpretable Machine Learning, Harvard University Feb 2022–May 2022
Spring Research Course. Advisors: Weiwei Pan, Yaniv Yacoby

- Investigated how non-identifiability in additive models can cause misleading model interpretations in the healthcare domain.
- Characterized a particular form of non-identifiability that arises when generalized additive models are trained on data with interaction effects.

Differential Privacy and Fairness, Microsoft — Sept 2021–Dec 2021
IACS Capstone Project. Advisors: Joshua Allen, Chris Tanner

- Led a collaboration among graduate students and Microsoft researchers to understand the fairness impact of training ML models on differentially private synthetic data.
- Proposed a simple pre-processing technique to synthesize data that promote more fair model predictions.

Epidemiological Modeling, Williams College — Feb 2020
Senior Mathematics Colloquium. Advisor: Julie Blackwood

- Applied compartmental models to early COVID-19 data published by the Chinese National Health Commission to estimate key disease parameters and simulate an outbreak on a college campus with a quarantine policy.

Professional Experience

Microsoft, Redmond, WA — Aug 2022–Present
Offensive Security Engineer, AI Red Team

- Leading red teaming of the Phi-3 language models including Phi-3-mini, small, medium and MoE (received the *CES Infinite Mindset Partnership Award* for June 2024).
- Researching gradient-based data exfiltration attacks against LLM-based Copilots with jailbreak filters.
- Testing a variety of generative AI models and products for harmful content and security vulnerabilities.
- Active contributor to the Python Risk Identification Tool for generative AI (PyRIT), an open-source framework that automates AI red teaming techniques.

Data Scientist

- Introduced a method to classify performance bugs and customer incidents using text embeddings (accepted to Microsoft's 2023 *Machine Learning and Data Science Conference*).
- Deployed an LLM-powered Azure web app that answers questions about internal documentation using retrieval augmented generation.
- Built a pipeline to detect and prioritize kernel-mode memory leaks across the Azure fleet (received a *Quality Stars* award for FY23 Q3).
- Trained ML models that help deployment teams assess the risk of Azure Host OS updates.

Marble — June 2020–Jan 2022
Co-Founder

- Led the development of an iOS mobile app that provides carbon footprint estimates for grocery products.
- Built Google Firebase backend with 150,000+ products scraped from supermarket websites.
- Accepted into the Harvard i-lab Venture Program for three consecutive semesters.

Teaching Experience

Graduate Teaching Fellow, Harvard University — Feb 2022–May 2022

- CS 109b: Advanced Topics in Data Science
- Prepared teaching materials and held office hours for students studying non-linear statistical methods and deep learning models, including CNNs, RNNs, LSTMs, autoencoders, GANs, and transformers.

Undergraduate Teaching Assistant, Williams College — 2017–2020

- CHIN 201: Intermediate Chinese I (Fall 2017)
- CHIN 202: Intermediate Chinese II (Spring 2018)
- CHIN 301: Upper-Intermediate Chinese I (Fall 2019)
- CHIN 302: Upper-Intermediate Chinese II (Spring 2020)
- In 1:1 sessions, met weekly with students for casual discussions to practice spoken language, review vocabulary, and learn grammar structures.

Service & Outreach

TEALS Program, Microsoft — August 2023–Present
Volunteer Teacher

- Delivering lectures and engaging with high school students to assist in teaching of AP Computer Science Principles at Global Impact Academy in Fairburn, GA.

IACS ComputeFest, Harvard University — Jan 2022
Volunteer Teaching Assistant

- Worked alongside professors to run workshop focused on teaching fundamental data science skills, including Python programming, probability theory, linear algebra, and statistics.

Honors & Awards

CES Infinite Mindset Partnership Award, Microsoft 2024
For safeguarding the Phi-3 language models as AI Red Team lead.

Quality Stars Award, Microsoft 2023
For building a novel memory leak detection pipeline for Azure.

Certificate of Distinction in Teaching, Harvard University 2022
Awarded based on student ratings (mean 4.67/5) for teaching of CS 109b.

IACS Student Scholarship, Harvard University 2021
Awarded to support data science thesis research at IACS ($20,000 award).

Goldberg Prize in Mathematics, Williams College 2020
Awarded to the graduating senior who delivers the best mathematics colloquium.

Linen Senior Prize in Chinese, Williams College 2020
Awarded to the top graduating Chinese major.

Putnam Competition, MAA 2019
Scored 18.

Carolyn Altes Scholarship, AWCA 2019
Awarded on the basis of academics and potential to contribute to society.

Linen Grant, Williams College 2017
Awarded on the basis of academics to support summer study in China.

Davis UWC Scholar, Davis United World College Scholars Program 2016
Awarded to recognize commitment to building cross-cultural understanding.

Class of '16 Student Speaker, UWCSEA East 2016
Elected by peers to deliver the Class of '16 graduation student address.

Skills & Interests

Programming: Python (NumPy, pandas, sklearn, TensorFLow, PyTorch), R, SQL, KQL, HTML/CSS, JavaScript

Tools/Platforms: Conda, Jupyter, Git, Docker, Kubernetes, Azure, AWS

Language: Working proficiency in written and spoken Chinese (Mandarin)

Interests: Running, rowing, writing (Medium blog), Rubik's cube solving (WCA profile)

References

Dr. **Pavlos Protopapas**

Harvard University
Email: pavlos@seas.harvard.edu

Dr. **Weiwei Pan**

Harvard University
Email: weiweipan@g.harvard.edu

Dr. **Mihai Stoiciu**

Williams College
Email: mstoiciu@williams.edu

Dr. **Julie Blackwood**

Williams College
Email: jcb5@williams.edu