Brad Cao

(301) 356-1069 | bcao4049@gmail.com | github.com/bradca0

Education

Columbia University New York, NY

B.A. in Mathematics & Computer Science *Expected May 2030*

- **Relevant Coursework:** Natural Language Processing, Probability Theory, Honors Mathematics A & B, Data Structures in Java
- **Activities:** Columbia AI Alignment Technical Fellowship, Columbia Alpha Partners (Global Markets Division), Blockchain @ Columbia (Investment Fund), Columbia Poker Club

Experience

Research Engineering Intern, Sampura Research · *Remote* Oct. 2026 – Present

- Onboard models and datasets onto Sampura's scalable-oversight evaluation infrastructure end to end, so new judge models and new eval sets can be scored against each other in production
- Vet ∼200 public AI safety and alignment datasets for threat-model fit, label provenance, ground-truth reliability, and contamination risk; request private or held-out sets from external research groups so judges are tested on data they cannot have memorized
- Deliver monthly summary reports and a final report on datasets reviewed, integrated, and outstanding

Research Fellow, Supervised Program for Alignment Research (SPAR) · *Remote* Sep. 2026 – Present

- Read, audit, and classify datasets for a 100+ dataset leaderboard that evaluates human+AI judge systems for scalable oversight across deception, jailbreak robustness, cultural bias, and computer-control mistakes
- Assess each dataset's ground truth, label quality, and threat-model fit before it enters the leaderboard; construct ground truth via sandwiching (AI-assisted non-experts checked against experts) where no expert labels exist
- Measure how judge robustness degrades under adversarial RL (reward hacking), the failure mode scalable oversight must resist

Data Science Intern, Mercor · *San Francisco, CA* Apr. 2026 – Sep. 2026

- Designed 10+ rubric-based scoring taxonomies that turn expert human judgment into consistent reward signal for LLMs (RLHF annotation schemas), the human-feedback layer alignment training depends on
- Wrote pass/fail criteria with worked examples plus ambiguous and conflicting edge cases so reviewers score outputs identically
- Built QA workflows and audited labeled data to raise inter-annotator agreement; resolved disagreements, refined guidelines

Research Assistant, Duke University · *Durham, NC* Apr. 2026 – Sep. 2026

- Designed and built the full data pipeline for a cable-news disparagement study, owning every technical and AI component, including LLM-based transcript classification validated against human-coded samples

Leadership Seminar Scholar, Notre Dame · *South Bend, IN* Jun. 2025 – Aug. 2025

- Won a full-ride scholarship to the "Inequality in America" seminar; produced original meta-research reviewing 15+ studies on the pharmaceutical industry's effect on public health policy and presented the findings at a symposium

Research Intern, University of Maryland, Baltimore · *Baltimore, MD* Jun. 2023 – Sep. 2025

- Trained logistic regression and random forest models on 10 years of FDA FAERS adverse-event reports to predict serious outcomes in ADHD patients on stimulants; best model 0.75 AUC / 88% recall, with age, sex, and drug type as the key risk factors
- First-authored 2 studies presented at the ASCPT Annual Meetings (2024, 2025); the 2025 poster won the Top Poster Award

Projects

Threshold | *Python, NumPy, SciPy, PyTorch* | github.com/bradca0/threshold Solo project

- Reproduced double descent with random-feature ridge regression on MNIST: test error peaks at 88.8% at the interpolation threshold (90% of it variance, by a bias–variance decomposition) and falls to 12.3% with a tuned ridge penalty
- Reproducible, tested Python package: 85 tests at 94% coverage with `mypy --strict`

MyCureHope | *Django, PostgreSQL, JavaScript, TailwindCSS* | mycurehope.org Co-developer

- Co-developed an open-source database of treatments, doctors, and hospitals used by 33,000+ users across 78 countries; personally built the search feature, own the technical development, and deployed the full stack on Nginx/Gunicorn/Ubuntu

Publications

Cao, B., Lee, S., Naga, A., Gobburu, J., Dunn, A. "Leveraging Supervised Machine Learning to Identify Predictors of Serious Clinical Outcome with Stimulant Treatment from Real-World Data." *ASCPT Annual Meeting*, 2025.

Cao, B., Gupta, P., Dunn, A. "Predictive Modeling of Tacrolimus Dose Requirements and Transplantation Outcomes in Lung Transplant Recipients on Systemic Azole Antifungals." *ASCPT Annual Meeting*, 2024.

Technical Skills & Interests

Alignment & Evals: Scalable Oversight, LLM-as-Judge Evaluation, Dataset Vetting and Curation, Sandwiching, Reward Hacking / Adversarial RL Robustness, RLHF Rubric Design, Mechanistic Interpretability (TransformerLens, GPT-2 Small IOI Circuit)
Languages & Tools: Python, Java, R, SQL (PostgreSQL), JavaScript; PyTorch, scikit-learn, NumPy, pandas, Django, Linux
Interests: Fencing (USA Fencing First Team All-American), Competitive Programming (USACO), History Olympiads, Open-Source