Brad Cao

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

Education

Columbia University — New York, NY
B.A. in Mathematics & Computer Science — *Expected May 2029*

Relevant Coursework: Data Structures in Java, Probability Theory, Natural Language Processing, Honors Mathematics A & B

Experience

Data Science Intern — Apr. 2026 – Present
Mercor — *San Francisco, CA*

- Engineered evaluation frameworks that formalize subject-matter expertise into RLHF-compatible annotation schemas for LLM fine-tuning pipelines
- Designed rubric-based scoring taxonomies and QA workflows to raise inter-annotator agreement; audited quality and flagged edge cases to refine labeling guidelines

Research Assistant — Apr. 2026 – Present
Duke University — *Durham, NC*

- Trained and developed LLMs to identify and quantify inter-network disparagement in televised news media, building NLP pipelines to ingest, segment, and label large transcript volumes
- Benchmarked prompt-engineered and fine-tuned classifiers against human-coded samples with Dr. Philip Napoli, translating media-studies frameworks into quantifiable schemas

Research Intern — Jun. 2025 – Aug. 2025
University of Notre Dame — *South Bend, IN*

- Awarded a full-ride scholarship for original quantitative research on the pharmaceutical industry's impact on public health policy
- Applied R-based regression and visualization to link industry practices to poverty outcomes; presented at a symposium

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

- First-authored a supervised ML model with Prof. Allison Dunn predicting ADHD-related patient safety risks from real-world data
- First-authored a study modeling tacrolimus dose requirements and transplant outcomes in lung transplant recipients on systemic azole antifungals
- Presented at the ASCPT Annual Meetings (2024, 2025); earned the Top Poster Award (May 2025)

Projects

Lookback | *Python, PyTorch, TransformerLens* | github.com/bradca0/induction-heads

- Located the attention heads behind in-context learning in GPT-2 Small: scored all 144 heads on repeated random-token sequences and found 16 induction heads (top head L5H5, score 0.94), recovering all five reported in the literature
- Proved they are causal by mean-ablating the top 4 via `hook_z`: second-copy loss rose from 0.23 to 7.83 nats (95% CI 7.66–8.00) while 4 matched-layer random heads left it at 0.27 ± 0.08 ($t = 86$); aggregate ICL score fell only 7.8% vs. 4.8% for controls ($t = 1.19$)
- Verified the two-step circuit (ablating previous-token head L4H11 cut the top induction score 29%); replicated in GPT-2 Medium, Pythia-160M, GPT-Neo-125M; induction heads and ICL emerge together in Pythia-70m at steps 512–1,000; 87 tests, `mypy --strict`

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

- Reproduced double descent on MNIST with a random-feature model (nested random ReLU features, min-norm least squares via SVD/Gram decompositions in float64; $n = 2{,}000$, 5 seeds): test error peaks at $p/n = 1$ (88.8% vs. 24.3% at $p = 5n$; 20% label noise)
- Explained the peak: $\|\beta\|_2$ and the smallest singular value of Φ are extreme at $p = n$ ($\|\beta\|_2$: 811 vs. 2.6 at $p = 5n$); a bias–variance split over 10 draws puts 90% of peak error in variance; tuned ridge cuts the peak from 88.8% to 12.3%
- Showed more data can hurt (fixed $p = 2{,}000$: test error 34.8% $\to$ 89.1% as n grows 250 $\to$ 2,000) and replicated peak and mechanism on Fashion-MNIST; a width-swept MLP showed no peak, reported as a negative result; 85 tests, 94% coverage, `mypy --strict`

MyCureHope | *Django, PostgreSQL, JavaScript, TailwindCSS* | mycurehope.org

- Co-developed an open-source patient database of treatments, doctors, and hospitals serving 33,000+ users across 78 countries; own search build and technical development
- Built and deployed the full-stack platform (Django, PostgreSQL, JavaScript, TailwindCSS) 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

Languages: Python, JavaScript, Java, SQL (PostgreSQL), R, HTML/CSS
Frameworks & Tools: PyTorch, TransformerLens, NumPy, pandas, scikit-learn, Django, TailwindCSS, Nginx, Gunicorn, Ubuntu/Linux
ML & Data: Mechanistic interpretability, LLM fine-tuning, RLHF annotation, supervised ML, data analysis
Interests: Fencing (USA Fencing First Team All-American), Competitive Programming (USACO), History Olympiads, Open-Source