Coby L. Kassner

cobylk.io | linkedin.com/in/cobylk
kassner@cobylk.io | +1 (720) 551-3481

Student researcher with broad interests in AI safety and mechanistic interpretability.

Experience

Research Fellow February–May 2025
Supervised Program for Alignment Research (SPAR)

- Researched neural networks that are inherently interpretable, mentored by Dr. Ronak Mehta
- Measured viability and interpretability of simplex-constrained MLP and Transformer variants

Student Researcher 2024–Present
Julia Student Research Group

- Headed project to extract synthetic training data from a fine-tuned Llama 3.1 8B instance
- Utilized contrastive activation addition to steer model outputs towards memorized examples
- Achieved ∼2x baseline success rate, placing 7th in the LLM Privacy Challenge, Red Team, at NeurIPS 2024

Student Researcher Summer 2024
Association of Students for Research in Artificial Intelligence

- Led project in natural language processing to understand dis/misinformation in the context of LLMs
- Benchmarked LLM fact-checking performance across 5 languages and several prompting techniques
- Co-first author publication in NLP4PI workshop at EMNLP 2024

Vice President, Outreach 2023–Present
International Research Olympiad (IRO)

- Directed program to start over 320 research clubs in secondary schools across 40 countries and 6 continents
- Collaborated with leadership team to coordinate over 50 student volunteers, negotiate over $15,000 in sponsorships to fund research clubs, and staff and organize in-person finals event in Cambridge, MA

Technical Skills

Research: LLM fine-tuning and inference to 70B scale, small-model training (GPT-2), custom transformer variants, activation steering, PINNs/PINOs, neuroevolution (NEAT, Hyper-NEAT, CPPNs)
Libraries: Transformers, PyTorch, JAX, Scikit-Learn, Pandas, NumPy, Transformer Lens
Languages: Python (6 years), C++ (3 years)

Education

Statistics and Data Science, B.S. 2025–2029
Yale College

Computer Science, A.S. and Mathematics, A.S. 2021–2025
Arapahoe Community College (concurrent enrollment)
Coursework: Multivariate Calculus, Linear Algebra, Computer Science II (C++; Data Structures, Algorithms, Object-Oriented Programming), Computer Architecture and Assembly, Discrete Structures, Calculus-based Physics

High School Diploma 2021–2025
Colorado Early Colleges Douglas County North
Class rank: 4/209