

# Coby L. Kassner

[cobyllk.io](https://cobyllk.io) | [linkedin.com/in/cobyllk](https://linkedin.com/in/cobyllk)  
kassner@cobyllk.io | +1 (720) 551-3481

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

## Experience

---

### Research Fellow

February 2025–Present

Supervised Program for Alignment Research

- Researching models that are inherently more interpretable, mentored by Dr. Ronak Mehta
- Experimenting with and measuring interpretability of simplex-constrained neural networks

### Student Researcher

2024–Present

Julia Student Research Group

- Headed project to extract synthetic training data examples from a fine-tuned Llama 3.1 8B instance
- Utilized contrastive activation addition to steer model outputs towards memorized examples
- Achieved  $\sim 2\times$  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

### Vice President, Outreach

2023–Present

International Research Olympiad

- Directed program to start over 280 research clubs in secondary schools across 35 countries and 5 continents
- Collaborated with leadership team to coordinate over 50 student volunteers and negotiate over \$15,000 in sponsorships to fund research clubs and in-person finals

## Education

---

### Computer Science, B.S.

2025–2029

Yale College

### Computer Science, Mathematics, $2\times$ A.S.

2021–2025

Arapahoe Community College

### High School Diploma

2021–2025

Colorado Early Colleges Douglas County North

## Technical Skills

---

**Research Experience:** Steering/activation engineering with LLMs, Physics-informed ML (PINNs, Fourier Features, PINOs), Genetic Algorithms (NEAT, Hyper-NEAT, CPPNs)

**Libraries:** Transformers, PyTorch, JAX, Scikit-Learn, Pandas, NumPy, Transformer Lens

**Languages:** Python, C++, SQL