Abraham Yeung

Stanford, CA · ayeung16@stanford.edu · +1 (650) 509-9458 · GitHub · LinkedIn

Education

Stanford University *Sept 2024 – 2028*
B.S. in Mathematics & Computer Science (GPA: 4.05/4.0) *Stanford, CA*

- CS / ML: Language Modeling from Scratch (CS 336), Deep Reinforcement Learning (CS 224R), Machine Learning (CS 229), Mining Massive Datasets (CS 246), Randomized Algorithms (CS 265)
- Math / Stats: Probability Theory, Stochastic Processes & Measure Theory (MATH 136), Statistical Inference, Real Analysis

Eton College *Sept 2019 – June 2024*
King's Scholar, Valedictorian *Windsor, UK*

Research & Projects

Activation Cache Compression for Sparse Autoencoder Training *2026*
Independent Research; workshop paper under review *Stanford, CA*

- Measured how aggressively cached LLM activations can be compressed before retrained sparse autoencoders diverge: ∼170 GPU-hours of pre-registered experiments, every published number regenerating from committed artifacts

Alembic: Auditing Cue-Injection Faithfulness Metrics *2026*
Independent Research; workshop paper under review *Stanford, CA*

- Showed the leading chain-of-thought faithfulness metric inherits its conclusion from size-dependent error terms: a content-free placebo reproduces up to 85% of measured cue influence at the smallest scale, replicating across runs

Weak-to-Strong Reasoning and Reasoning-Trace Monitorability *2026*
Independent Research (CS 338) *Stanford, CA*

- Designed a pre-registered study with explicit validity gates and stopping rules, testing how a training procedure degrades the reliability of model reasoning traces
- Built the pipeline on Modal GPU with LoRA fine-tuning and an independent model judge; established significance with bootstrap confidence intervals

Probing Internal Self-Knowledge in Reasoning Models *2026*
Connected thread of the reasoning-trace research line above (CS 224R) *Stanford, CA*

- Trained linear probes on hidden states to predict rollout correctness: as a read-only best-of-16 selector, a probe captures 89% of available headroom (held-out AUROC 0.939 with trace length stratified out)
- Promoted to an RL reward, a probe of the same representation is catastrophic: probe score rises while true accuracy falls ∼31 points, so reading a representation and optimizing against it are different privileges

Experience

AI Safety Researcher Intern (Incoming, Part-Time) *Fall 2026*
Redwood Research *Berkeley, CA*

Undergraduate Researcher *Jun 2025 – Sep 2025*
Chiu Lab, Stanford University (Bioengineering REU Program) *Stanford, CA*

- Built CNN and Vision Transformer models (CryoViT) to automate cell-structure segmentation and annotation in noisy 3D cryo-EM imaging data

Software Engineer Intern *Jun 2026 – Sep 2026*
Databricks (Data Platform) *San Francisco, CA*

Awards

- **British Mathematical Olympiad Round 2** – Top 50 nationally (2x); **UK Chemistry Olympiad** Gold Medal (3x), IChO team reserve
- **World Science Scholars** – 1 of 48 globally; **Rabi Scholar**, Columbia University – top 10 scientific admits nationally

Technical Skills

- **ML / Data:** PyTorch, NumPy, Pandas, SciPy, scikit-learn, Matplotlib, R
- **Concepts:** RL post-training (GRPO, DPO, RLHF), interpretability / probing, Monte Carlo methods, time series
- **Systems & Programming:** Python, C++, SQL, Bash, Modal, vLLM, LoRA, Git, Linux, Docker, CI/CD, profiling