Anthony Nguyen

408-228-2849 | anthony.nguyen@berkeley.edu | linkedin.com/in/antharny | github.com/anthnguyen

EDUCATION

University of California, Berkeley — Berkeley, CA
B.A. Mathematics — *Expected May 2029*

De Anza College — Cupertino, CA
A.S. Mathematics, GPA: 4.0 — *Sep. 2025 – Jun. 2026*

EXPERIENCE

SPAR Research (F26) — Remote
Research Fellow — *Sep. 2026 – Present*

- Investigating whether interventions validated on model organisms transfer to naturally misaligned models; mentored by Jeanne Salle and Sohaib Imran
- Identifying which training-time hyperparameters used to build model organisms make interventions transfer

Berkeley AI Safety Initiative (BASIS) — Berkeley, CA
Member, Paper Replication Group — *Oct. 2026 – Present*

- Member of the first paper replication group; replicating AI safety papers and extending their findings
- Releasing open-source code and publishing write-ups on LessWrong

Algoverse (S25) — Remote
Research Fellow — *Jun. 2025 – Jul. 2026*

- Trained linear probes on alignment-faking model organisms; mentored by Yuqi Sun and Josh Ying

RESEARCH & ENGINEERING PROJECTS

How Post-Training Dataset Construction Affects Model Organism Transfer | Code — *SPAR Research, 2026*

- Conditional, less leaky model organisms transferred better, removing 18 more points of natural sycophancy on Llama ($p = 0.003$) and 5 more on Qwen ($p < 0.001$)
- Ablating sycophancy directions from 14 model organisms each on Llama-3.1-8B and Qwen2.5-7B removed 25–69% of the parent's natural sycophancy, beating random directions
- Mixing in benign data did not change transferability ($p = 0.15, 0.60$); off-policy data lowered it in 9/10 pairs
- Found Llama's contrarian model organisms hid sycophancy rather than removing it (0.35 on medical vs 0.91 in-domain); self-ablation strength predicted transfer ($\rho = 0.82$)
- Standard pushback understated Qwen's natural sycophancy (0.10); a pre-registered rule raised it to 0.53

Sample-Level White-Box Detection of Alignment Faking Model Organisms | Code — *ICML 2026 Workshops*

- Built the first alignment-faking-specific synthetic training set (600 samples, 6 conditions) from scenario templates; probes trained on the dataset beat prior lie and deception probes by 28% AUROC, 33% AUPRC
- Alignment-faking research lacked per-sample ground truth, designed an LLM-as-judge flowchart for labeling model rollouts, scaling to 3,120 rollouts across 32 conditions
- Found model organisms need model-specific auditing: probe directions drifted 0.15–0.40 cosine in fine-tuning
- Pulled Llama-3-70B residual-stream activations from two alignment-faking model organisms (LoRA fine-tunes) via nnsight at layers 33/37/40 of 80, then swept 96 probe configs per model under 5-fold CV

Conditionally Incorrect Model Organism | Code — *Independent Research, 2026*

- Fine-tuned Qwen2.5-7B-Instruct into a conditionally incorrect model organism that gives false animal-size answers only when a token appears; trained 15 rank-8 LoRA adapters across 5 experimental/control families
- Probed all 29 residual positions to test how the fine-tune changed its animal-size encoding; new probes reached 99.6% AUROC versus 71.9% accuracy for the base-model probe

SELECTED PUBLICATIONS

A. Nguyen et al. "Sample-Level White-Box Detection of Alignment Faking." **ICML** Mechanistic Interpretability, 2026.
A. Nguyen et al. "Sample-Level White-Box Auditing of Alignment Faking." **ICML** TAIGR Workshop, 2026.

TECHNICAL SKILLS

Languages & Systems: Python, C++, Bash, Git, Linux, LaTeX, SLURM
ML & Data: PyTorch, Hugging Face Transformers/Datasets, nnsight, PEFT/LoRA, scikit-learn, NumPy, pandas
Interpretability & Stats: residual-stream capture, activation patching, linear probes, bootstrap CIs, TOST, k-fold CV