Tony Gao

Machine learning & interpretability research

tonygao0112358@gmail.com | New York, NY
github.com/tgao09 | linkedin.com/in/tonygao09

Education

Horace Mann School Expected June 2027
High school diploma
Calculus with Theory (Honors); Math Seminar (discrete math, complex geometry, multivariable calculus)

Carnegie Mellon University Summer 2026
21-241 Matrices and Linear Transformations (A); 21-127 Concepts of Mathematics (A)

Research

When Probes Fail to Steer: Decodability and Causal Control in Closed-Loop RL 2026
Sole author · NeurIPS 2026 Workshop on Interpreting Agent Behavior (accepted)

- Built a closed-loop protocol testing whether features that *predict* a frozen Rocket League agent's behavior can also *control* it, scoring live activation edits on target lift, policy drift and side effects.
- Trained a 1,024-feature sparse autoencoder on ∼540K activations (0.969 explained variance) and future-event probes at 0.89–0.96 AUROC.
- Of 624 edits screened, most failed to steer; one SAE feature raised dribble rate from 42.8 to 65.1 per 1k steps on 11 of 12 held-out seeds (KL 0.0047).

CIPHER: Cryptographic Insecurity Profiling via Hybrid Evaluation of Responses 2025–2026
Co-first author · equal contribution, arXiv:2602.01438

- Co-built a benchmark of 450 insecure / neutral / secure prompts and a 58-type crypto vulnerability taxonomy; evaluated 7 LLMs incl. GPT-5.1 and Claude Sonnet 4.5.
- Built AST and regex detectors and an LLM judge with line-level evidence, at 85.65% accuracy ($\kappa = 0.77$) against a human audit.
- Found flaws in 56–89% of generated code; explicit security prompts only cut the pooled rate from 76.9% to 71.7%.

Enhancing Viewer Experience on YouTube: ML for Assessing Tutorial Video Quality 2023–2024
Co-author · National High School Journal of Science, March 2024

- Rated 50 tutorials from engagement ratios and VADER comment sentiment, the top predictor ($r = 0.85$).
- Compared seven models; a random forest reached 0.06 test MSE versus 0.195 for linear regression.

Projects & Competitions

IMC Prosperity 4 Global Trading Challenge April 2026
Team lead · IMC's global trading-simulation challenge

- Led our team to the **top 5% of 18,803 teams** worldwide over the 16-day competition.
- Wrote Python trading algorithms and solved manual rounds spanning market making, options trading and risk analysis.

Market Analysis 2025–2026

- Forecast stock returns and volatility with gradient boosting, Gaussian processes, LSTMs and hidden Markov models.
- Trained a PPO agent for multi-asset portfolio allocation, evaluated on Sharpe ratio and max drawdown.
- Wrote a reusable backtester with chronological train/test splits and data-leakage safeguards.

Leadership & Honors

Citizens of Tomorrow 2022–present
President of Program Planning · free English classes for immigrant learners

- Co-lead 15 classes serving 200 learners and coordinate 100 student volunteers; helped build the curriculum and logged 260+ teaching hours.

Honors: National Merit Semifinalist · USACO Silver · President's Volunteer Service Award (Bronze)
Music: Piano for 10+ years; performed at Carnegie Hall's Weill Recital Hall

Skills

Languages	Python, Java, JavaScript
Libraries	PyTorch, TransformerLens, scikit-learn, pandas, Matplotlib
Methods	Sparse autoencoders, linear probes, activation steering, LLM evaluation, RL (PPO), forecasting