# Hong Kiat, Tan

maxtanhk@math.ucla.edu | maxtanhk.com | github.com/HK-Tan


## Education

### University of California, Los Angeles (UCLA)

- **Ph.D.** Candidate, Mathematics. Aug 2021 – Dec 2026
- Advisor: Andrea L. Bertozzi. (GPA: 3.98/4.00)

### National University of Singapore (NUS)

- **Bachelor of Science (Honours)** in Applied Mathematics with Highest Distinction. Aug 2017 – May 2021
- Second Major in Physics and Minor in Statistics. (GPA: 4.97/5.00)
- *Ho Family Prize* – Top graduating student in Applied Mathematics, with 28 A+'s in Math/Physics/Statistics courses.

## AI Safety Research Experience

*First or co-first author.

*Mentor,* Supervised Program for Alignment Research Sep 2026 – Present

- Mentoring two projects: model organisms and interpretability for scheming, and modeling and detecting unspoken collusion (Schelling point, tacit, and acausal), with D. Williams-King and L. Le.

*Iliad Intensive Participant* Oct 2026

*Research Collaborator,* Lida Safety Jun 2026 – Present

- Introduced Activation Flow, a training-free differential-geometric method that manufactures steering activations from answer labels alone, recovering 92% of the sandbagging gap (SFT: 96%) with 3,000× fewer parameters than SFT. (NeurIPS 2026 NeurReps Workshop*; full version: arXiv:2609.32530*.)
- Proposed a causal model of sandbagging in the residual stream to explain why steering works, and validated it across 11 LLMs (1.5–32B) and 4 lock types (prompting, fine-tuning, RL, circuit-breaking). (arXiv:2608.29461*, arXiv:2608.29458.)

*Research Fellow,* Supervised Program for Alignment Research Feb 2026 – Jun 2026

- Developed Circuit Oracle, a training-free multi-agent system that answers safety-relevant queries over LLM circuits through composable skills and custom interpretability tools. (ICML CompLearn* & Mech. Interp.* Workshops, 2026.)

## Industrial/Work Experiences

*Applied Scientist Intern,* Amazon Search Jun 2026 – Sep 2026

- Developing statistically principled RL methods and ranking theory to post-train LLM search rankers (paper in progress).
- Ran ∼120,000 H100-hours of multi-node training and inference on verl/vLLM.

*Quantitative Research Intern,* WorldQuant Intraday Team Sep 2025 – Dec 2025

*Data Scientist Intern,* Amazon Search Data Science and Economics Jun 2025 – Sep 2025

- Built a framework orchestrating multi-agent LLMs with $\ell^0$-changepoint detection for 78 search metrics in 17 locales to generate interpretable economic insights via AWS AgentCore. (ICLR TSALM Workshop 2026*.)

## Academic Experiences

*Graduate Research Assistant,* UCLA 2022 – 2026

- Proved generic structural stability of Riemann solutions for hyperbolic conservation laws by generalizing Thom's parametric transversality theorem to a foliated setting and introducing sequential transversality for causally chained submanifolds, the same geometric toolkit behind Activation Flow. (SIAM J. Math. Anal., 2026*, arXiv:2609.28714*.)
- Developed a modified vector autoregressive framework using double machine learning and causal inference for confounding-adjusted lag detection in time series, with EconML, TabPFN, and PyTorch in Python. (arXiv:2610.05618*.)
- Architected an object-oriented Python framework coupling game-theoretic models and PDE solvers on directed graphs, optimizing evacuation routing on a non-smooth loss. (Discrete and Continuous Dynamical Systems, 2027*.)

*Graduate Teaching Assistant,* UCLA 2021 – 2025

- Developed 786 pages of instructional materials across 10 quarters for advanced mathematics courses (evaluation 8.6/9.0).

*Undergraduate Research Assistant,* NUS & UNC Chapel Hill 2019 – 2021

- Built a numerical scheme with hypothesis testing and regression in R for quantum field theory. (Phys. Rev. D, 2022*.)
- Co-authored a 148-page paper on a fundamental conjecture in mathematical general relativity. (arXiv:2402.16250*, 2024.)
- Designed Bayesian hierarchical models for astrophysical data with MCMC samplers in R at UNC. (Astrophys. J., 2020.)

## Technical Skills & Professional Activities

- *Programming Languages*: Python, C++ (Intermediate), PostgreSQL, R, LaTeX.
- *Certification*: BlueDot Technical AI Safety Course.
- *Cloud & DevOps*: AWS (AgentCore, Lambda, ECS/ECR, Bedrock), GCP, Modal, RunPod, Docker, Linux (SLURM), Git.
- *Libraries/Frameworks*: PyTorch, vLLM, verl, SGLang, Hugging Face, wandb, TabPFN, EconML, sklearn.
- *Service*: Reviewer for AISTATS 2026/2027, UAI 2026, NeurIPS 2026, IASEAI 2027, ICML/NeurIPS Workshops 2026.