Daniel (Chee Hian) Tan

Personal site: https://dtch1997.github.io/
Email: dtch009@gmail.com

About Me

I am a current PhD student at University College London (UCL). My current research direction focuses on scalable ways to interpret deep neural networks, with a focus on the architectures used in foundation models. I believe that better interpretability for foundation models, both in language and in other domains, will lead to safe and powerful agents which can be deployed in the real world for societal benefit.

Education

SEP 2017 - SEP 2021
Stanford University – *B. Sc. Mathematical and Computational Sciences*

SEP 2022 - PRESENT
University College London – *M. Phil/PhD. Computer Science*

Publications

Tan and Chanin et al. 'Analyzing the Reliability and Generalization of Steering Vectors'. Accepted to ICML 2024 Workshop on Mechanistic Interpretability. [Arxiv upcoming]

S. Huang *et al.*, 'Open RL Benchmark: Comprehensive Tracked Experiments for Reinforcement Learning'. arXiv, Feb. 05, 2024. doi: 10.48550/arXiv.2402.03046.

Tan, D.C., Acero, F., McCarthy, R., Kanoulas, D., & Li, Z. (2023). Value Functions are Control Barrier Functions: Verification of Safe Policies using Control Theory. *ArXiv, abs/2306.04026*.

Tan, D.C.*, Zhang, J.*, Chuah, M.I., & Li, Z. (2023). Perceptive Locomotion with Controllable Pace and Natural Gait Transitions Over Uneven Terrains. *ArXiv, abs/2301.10894*.

Darici, E., Rasmussen, N., Tan, D.C. Ranjani, J.J., Xiao, J., Chaudhari, G.R., Rajput, A., Govindan, P., Yamaura, M., Gomezjurado, L., Khanzada, A., & Pilanci, M. (2022). Using Deep Learning with Large Aggregated Datasets for COVID-19 Classification from Cough. *ArXiv, abs/2201.01669*.

Industry Experience

JUL 2021 - AUG 2022

Agency of Science, Technology, and Research, Singapore- *Research Engineer*

AUG 2022 - OCT 2022

Virufy - *MLOps Tech Lead*

JUN 2019 - DEC 2019

GovTech, Singapore - *Software Engineering Intern*

JUN 2019 - SEP 2019

TripAdvisor, Boston - *Software Engineering Intern*

Teaching

University College London

- COMP0188 Deep Representations and Learning, 2022
- COMP0233 Research Software Engineering in Python, 2022
- COMP0016 Systems Engineering, 2023