Tung-Yu (Tony) Wu

tony10101105@gmail.com Google Scholar
LinkedIn Github

Research Interest

My research interests center on AI safety and interpretability. I am particularly interested in understanding how intelligence emerges and evolves in AI agents as they interact with the world, and in ensuring that such intelligence is aligned with and beneficial to humans.

Education

National Taiwan University (NTU) **2019 – 2024**
Bachelor: Major in Electrical Engineering, Double Major in Economics

Research Experience

Dr. Fazl Barez, University of Oxford **2025– Present**
Research Intern at Oxford *Oxford, UK*

- Work in progress: "ARIA: An Agent-Driven Research Environment for Automated Interpretability and AI Safety"
- Lucas Irwin, **Tung-Yu Wu**, Fazl Barez. "Token Taxes: mitigating AGI's economic risks," arXiv preprint arXiv:2603.04555, 2026.
- **Tung-Yu Wu**, Fazl Barez. "Query Circuits: Explaining How Language Models Answer User Prompts," arXiv preprint arXiv:2509.24808, 2025.
- Fazl Barez, **Tung-Yu Wu**, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, Adel Bibi, Robert Trager, Damiano Fornasiere, John Yan, Yanai Elazar, and Yoshua Bengio. "Chain-of-thought is not explainability," Preprint, alphaXiv:2025.02v2, 2025.

Professor Pei-Yu Lo, NTU **2025– 2025**
Research Assistant *Taipei, Taiwan*

- **Tung-Yu Wu** and Pei-Yu Lo, "U-shaped and Inverted-U Scaling behind Emergent Abilities of Large Language Models," in Proceedings of the International Conference on Learning Representations (ICLR), 2025.

Professor Tsui-Wei (Lily) Weng, UC San Diego **2024– 2024**
Visiting Student *San Diego, US*

- **Tung-Yu Wu**, Yu-Xiang Lin, and Tsui-Wei Weng, "AND: Audio Network Dissection for Interpreting Deep Acoustic Models," in Proceedings of the International Conference on Machine Learning (ICML), 2024.

Professor Yu-Chiang (Frank) Wang, NTU&NVIDIA **2020– 2024**
Research Student *Taipei, Taiwan*

- **Tung-Yu Wu**, Sheng-Yu Huang, and Yu-Chiang Wang, "Data-Efficient 3D Visual Grounding via Order-Aware Referring," IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2025.

Professor Hung-Yi Lee, NTU **2021– 2023**
Research Student *Taipei, Taiwan*

- **Tung-Yu, Wu**, Tsu-Yuan Hsu, Chen-An Li, Tzu-Han Lin, and Hung-yi Lee, "The efficacy of self-supervised speech models for audio representations," in Proceedings of HEAR: Holistic Evaluation of Audio Representations. NeurIPS'22 Competition Track, 2022. **Lightning Talk**.
- Tsu-Yuan Hsu, Chen-An Li, **Tung-Yu, Wu**, and Hung-yi Lee, "Model Extraction Attack against Self-supervised Speech Models," arXiv preprint arXiv:2211.16044, 2022.

Personal Research Project **2021– 2021**
Side Project *Taipei, Taiwan*

- **Tung-Yu, Wu** and You-Ting Wang, "Locally interpretable one-class anomaly detection for credit card fraud detection," in Proceedings of the IEEE International Conference on Technologies and Applications of Artificial Intelligence (TAAI), 2021. **Best Paper Award**.

Internships

PranaQ **2022– 2022**
Medical AI Research Intern *Taipei, Taiwan*

- Cooperate with doctors to build AI models for automatic Photoplethysmography (PPG) signal quality assessment, supervised by Professor Hau-Tieng Wu at Courant Institute of Mathematical Sciences, NYU

VIE technologies **2020– 2021**
Deep Learning and Computer Vision Research Intern *New Taipei City, Taiwan*

- Leverage adversarial training to improve the performance of 2D object detection models.
- Help develop an autoML platform where clients can customize their AI models easily.

Awards

Oral Presentation | NeurIPS'24 2nd Workshop on Attributing Model Behavior at Scale (ATTRIB) **2024**

Infineon Enterprise Award with project exhibited at COMPUTEX 2022 | 2022 MakeNTU Hackathon **2022**

Lightning Talk in NeurIPS'21 Competition Session | NeurIPS'21 HEAR Challenge **2021**

Best Paper Award | 2021 TAAI **2021**

Infineon Enterprise Award | 2021 MakeNTU Hackathon **2021**

Grant Award | 2020 MakeNTU Hackathon **2020**

2nd place in the group of VIA technologies | 2020 HackMeiChu **2020**

Merit Award | 13th NTU Physics Creative Experiment Design Competition **2020**

Taiwan Representative | 2019 International Swiss Talent Forum **2019**