Wei Jie Yeo

AI Safety Researcher

Singapore

yeoweijie1995@gmail.com | https://wj210.github.io/ | https://www.linkedin.com/in/yeoweijie210/

Summary

PhD researcher with extensive experience in AI interpretability and safety. My research focuses on understanding complex and emergent behaviors in advanced AI systems and leveraging these insights to develop methods that provide stronger guarantees of their safety and reliability.

Education

Nanyang Technological University *Singapore*

PhD in Computer Science *2022 - 2026*

- Advisor: Prof. Erik Cambria

Nanyang Technological University *Singapore*

Bachelor in Mechanical Engineering with Honours (Distinction) *2017 - 2020*

Professional Experience

03/26 - 09/26	**Machine Learning Engineer**, TikTok
07/25 - 12/25	**AI Safety Research Intern**, Huawei International Pte Ltd
2025 Spring	**Mechanistic Interpretability MATS Scholar**, Advised by Neel Nanda
2020-2022	**Equipment Engineer**, Qualcomm

Publications

Wei Jie Yeo, Ranjan Satapathy, Goh Siow Mong, Erik Cambria. How Interpretable are Reasoning Explanations from Prompting Large Language Models? NAACL, 2024

Wei Jie Yeo, Ranjan Satapathy, Erik Cambria. Plausible Extractive Rationalization through Semi-Supervised Entailment Signal. ACL, 2024

Wei Jie Yeo, Teddy Ferdinan, Przemyslaw Kazienko, Ranjan Satapathy, Erik Cambria. Self-training Large Language Models through Knowledge Detection. EMNLP, 2024

Wei Jie Yeo, Wihan Van Der Heever, Rui Mao, Erik Cambria, Ranjan Satapathy, Gianmarco Mengaldo. A comprehensive review on financial explainable AI. AIRE Journal, 2025

Wei Jie Yeo, Nirmalendu Prakash, Clement Neo, Roy Ka-Wei Lee, Ranjan Satapathy, Erik Cambria. Understanding Refusal in Language Models with Sparse Autoencoders. EMNLP, 2025

Wei Jie Yeo, Ranjan Satapathy, Erik Cambria. Towards Faithful Natural Language Explanations: A Study Using Activation Patching in Large Language Models. EMNLP, 2025

Nirmalendu Prakash, **Wei Jie Yeo**, Amir Abdullah, Ranjan Satapathy, Erik Cambria, Roy Ka-Wei Lee. Beyond I'm Sorry, I Can't: Dissecting Large-Language-Model Refusal, AAAI, 2026

Wei Jie Yeo, Rui Mao, Moloud Abdar, Ranjan Satapathy, Erik Cambria. Debiasing CLIP: Interpreting and Correcting Bias in Attention Heads. ICML, 2026

SKills

AI/ML Programming

DeepLearning	**Torch, Numpy, Pandas, TRL**,
NLP	**Transformers, vllm**,
Interpretability	**Transformerlens, NNsight, SAElens**,

Research Experience

CoT Monitoring: Experience in implementing various forms of monitoring on Chain-of-Thought reasoning (CoT), including backtracking behaviors and evaluating faithfulness, plausibility and simulatability.

Mechanistic Interpretability (MI): Experience with implementing MI tools such as logit attribution, activation patching, circuit tracing, feature attribution with Sparse Autoencoders (SAES) and linear probing.

Behavior steering: Experience with utilizing internal activations to steer a model towards targeted behaviors such as backtracking, harmful compliance or bias reduction.

AI Robustness: Experience with using interpretability tools to mitigate spurious bias in vision models such as CLIP, and presenting an explanation for why such biases exist.

AI Safety: Experience with evaluating LLMs across jailbreaking, prompt injection, misalignment, etc. Additionally, equipped with research experience in designing defense frameworks against adversarial attacks. Other relevant experience includes detecting hallucination and improving factuality in LLMs.

AI Alignment: Research experiment in conducting misalignment evaluations across frontier models, including deception, eval-awareness and value misalignment.

Agentic AI: Research experience in optimizing for safety vs utility retention in LLM agents and multi-modal GUI agents.

LLM training: Experience with conducting traditional fine-tuning: SFT, DPO, RL on open-sourced models under full parameter and low-resource (LoRA) settings.