Ivan Lin

Sydney, Australia | ivanlin1818@gmail.com | https://www.ivanl.in | https://github.com/ivanlin9

EDUCATION

University of New South Wales — Bachelor's studies, Science/Mathematics pathway
Expected graduation: 2029
Relevant interests: mathematics, machine learning, AI alignment, mechanistic interpretability, AI control, and compute verification.

GRANTS & AWARDS

Emergent Ventures Grant, 2024 — Received support for independent research in AI safety and interpretability.

Defensive Acceleration Hackathon — 4th Place, 2025 — Apart Research × BlueDot Impact
Placed 4th out of 81 submissions for a project on privacy-preserving model deployment.

RESEARCH & TECHNICAL PROJECTS

Automating Privacy-Preserving Model Deployment — Def/Acc Hackathon, 2025
Built a system for automatically identifying and replacing homomorphic-encryption-unfriendly operations in neural network inference pipelines. Used PyTorch graph analysis and polynomial approximations to make model components more compatible with encrypted inference settings.

AI Control and Perturbation-Based Resampling — Independent research direction, 2026
Designed and attempted to implement experiments testing whether controlled perturbations to model activations, embeddings, or weights could improve detection of malicious behaviour in Ctrl-Z-style AI control protocols. The project compared ordinary stochastic resampling against perturbation-based resampling, with planned evaluation on safety/usefulness tradeoffs and attack discovery rates.

Estimating Local Learning Coefficients to Probe Loss Landscape Robustness — AI Safety × Physics Grand Challenge Hackathon, 2025
Implemented experiments inspired by Singular Learning Theory to estimate local learning coefficients and compare them against loss sensitivity under random weight perturbations. Explored whether geometric properties of neural network loss landscapes could provide useful signals about robustness or developmental structure.

Automated Manhwa-to-Video Generation Pipeline — Independent project, 2024
Built a pipeline for converting raw manhwa panels into narrated video outputs, including panel processing, sequencing, narration, and automation of editing steps.

TECHNICAL SKILLS

Programming: Python, shell/Linux, PyTorch, Hugging Face Transformers
ML/AI tooling: TransformerLens, PEFT/LoRA, nnsight, model activation extraction, evaluation pipelines
Research: experimental design, literature review, technical writing, interpretability experiments, safety evaluation
Infrastructure: GPU experimentation, Linux/VPS usage, basic deployment workflows