Hieu Minh "Jord" Nguyen

jordnguyen43@gmail.com | Google Scholar | GitHub | LessWrong

Research Experience

Constellation Institute Jan – Feb 2026
Visiting Researcher
- Led research on improving LLM introspective access and self-reports.
- Continuing research on activation monitoring and oversight

Soothcheck Oct 2025 – Jan 2026
Research Lead
- Leading research on latent space monitors to improve AI oversight (e.g., probes for AI debate)
- Collaborating with researchers from PIBBSS, Stanford University, and University of Bristol
- Supported by a research grant from Coefficient Giving

Trajectory Labs Oct – Dec 2025
Redteaming Contractor
- Red-teamed frontier AI agents
- Built RL environments for AI robustness training

Pivotal Research Feb – Apr 2025
Technical Governance Research Fellow
- Led research on measuring and monitoring evaluation awareness
- Used linear probes to detect when models recognise evaluation contexts
- Collaborated with researchers from Apollo Research, UCLA, and Waseda University

Apart Research Dec 2023 – Feb 2025
Research Fellow
- Led and published 3 research papers (ICLR, AAAI, TAIS)
- 1st place AI Security Evaluation Hackathon; 3rd place LLM Evaluations Hackathon

Publications

Probing and Steering Evaluation Awareness of Language Models [paper] *ICML 2025*
Technical AI Governance & Actionable Interpretability Workshops. First author.

Identifying Cooperative Personalities through Personality Steering [paper] *TAIS 2025*
Co-author.

DarkBench: Benchmarking Dark Patterns in Large Language Models [paper] *ICLR 2025*
Oral Spotlight. Co-first author.

Evaluating Precise Geolocation Inference Capabilities of VLMs [paper] *AAAI 2025*
Datasets and Evaluators of AI Safety Workshop. Co-first author.

Evaluating LLM-Based Agents in Mixed-Motive Scenarios [paper] *NeurIPS 2025*
Co-author.

Technical Writing

Here's 18 Applications of Deception Probes (Aug 2025) [post]
Can SAE steering reveal sandbagging? (Apr 2025) [post]
A Survey of Theory of Mind in Large Language Models (Feb 2025) [paper]
Shallow Review of Technical AI Safety 2025 (Dec 2025) [post]
Shallow Review of Technical AI Safety 2024 (Dec 2024) [post]

Selected Talks

NeurIPS 2024 – Concordia Workshop: Narrative steering for LLM cooperation [recording]

REAIM Summit 2024 – Invited by UN Office for Disarmament Affairs to present on catastrophic AI risks

ASEAN Regional Forum on Non-Proliferation and Disarmament 2025 – Invited to discuss challenges in AI safety and security.

EA Summit Vietnam 2025 – Introduced AI safety research and catastrophic risks to general audience

Selected Awards

1st prize, Sci-fAI Futures Youth Challenge (UNODA), 2024

1st place, AI Security Evaluation Hackathon (Apart Research), 2024

3rd place, ML Model Attribution Challenge (DEFCON 30 AI Village), 2022 – presented at IEEE SaTML 2023

Skills

Interpretability: TransformerLens, nnsight, linear probes, steering vectors
ML: PyTorch, HuggingFace, LLM APIs (OpenAI, Anthropic, Google), LLM finetuning
Infrastructure: Git, Linux, remote compute (Runpod, vast.ai), running distributed experiment jobs

Education

University of Science and Technology of Hanoi – BSc. Data Science (Class of 2027)
GPA: 3.7/4 | ACT: 36 | Academic Excellence Scholarship

AI Security Bootcamp (Aug 2025) – 4-week intensive on security fundamentals for AI systems.
Cryptography, network security, penetration testing, reverse engineering, adversarial ML, LLM security.

Non-trivial Fellowship (Sep – Dec 2023)