Last Updated on 14th April 2025

Jonathan Ng

jonathan.ng1@gmail.com | @derpy
H1B1 Eligible | +65 83995537

EDUCATION

NUS (NATIONAL UNIVERSITY OF SINGAPORE)

BCOMP IN INFORMATION SECURITY
December 2021 | Singapore, SG

LINKS

Github:// **derpyplops**
LinkedIn:// **Jonathan Ng**

SKILLS

PROGRAMMING

Proficient:
Python • Scala • VueJS
JS • SQL • React

Familiar:
Java• Shell
Golang • C++ • C

TOOLS & FRAMEWORKS

Apollo GQL • ZIO • Play Framework
AWS • GRPC • Jest
AsciiDoc • Redux
express.js • Kafka • Docker

INTERESTS

Effective Altruism
Arete Fellowship (SG) Organizer (Fall '20, Spring '21)
NUS Chapter Lead

MISC

ARENA (arena.education)
• Implemented PPO, BERT, GPT (and more) from scratch.
• Contributed to TransformerLens, an open source library for interpretability research.

PROFESSIONAL EXPERIENCE

SINGAPORE AI SAFETY INSTITUTE (AISI) | SPECIAL PROJECTS ASSOCIATE

Nov 2024 - Present

- Organising and conducting different programs such as the AI Safety Fellowship, with two tracks: Technical and Governance
- Started the first AI Safety programs in two universities in Singapore.

APART RESEARCH | RESEARCH ENGINEER (CONTRACTING)

Apr 2024 – November 2024

- First author on ***Catastrophic Cyber Capabilities Benchmark (3CB): Robustly Evaluating LLM Agent Cyber Offense Capabilities*** (AAAI DATASAFE Workshop '25)
- Designed and implemented 11 complex cybersecurity LLM agent evals
- Implemented multi-Docker agent evaluations in our agent LLM scaffolding
- Wrote and conducted experiments across 14 language models, across Anthropic, OpenAI, TogetherAI, Replicate

SURVEYOR | SOLE ENGINEER

Oct 2023 – Dec 2024

- Sole engineer building Surveyor, a webapp to build and deploy surveys. Backend, UX design, architecture and prompt engineering.
- Implemented data pipeline for robustly parsing surveys from LM outputs, inclusive of conducting experiments and writing tests
- Oversaw sub-contractor work on productionizing via containerization.
  > VueJS, TypeScript, Firestore DB, OpenAI API

CADENZA LABS | RESEARCH ENGINEER

Apr 2023 – Present

- Developed the EleutherAI/elk library, a white-box LLM interpretability library for training "truth-seeking" probes
- Coordinated and conducted large sweeps across multiple GPU clusters, conducted original experiments

SERI MATS | RESEARCH ENGINEER

Jan 2023 – March 2023 | Singapore

- Participated in seminars and workshops relating to AI Alignment, mentored by Dr Dan Hendrycks.
- Authored the ***Machiavelli*** paper (ICML23), created a data collection platform on Streamlit (repo) (paper).
- Worked with the trlx library to train LMs on an ethics dataset via RLHF.

LEADIQ | SOFTWARE ENGINEER (FULL-STACK)

Apr 2020 – June 2022 | Singapore

- Built Self-Serve Single Sign-On, a feature that automates SSO client setup, and cuts a 2-3 week long process down to a day. Won internal hackathon 1st place. After winning, we brought the code and made it production ready in a month.
- Built Slack Bot delivering Job Change Notifications.
  > Kafka, Scala, ZIO, Slack API
- Built Assign Sequences with Outreach.io API.
  > Vue, CSS, HTML, ApolloGQL, Scala