

ADITYA NAWAL

+91 8769844417 | ✉ adityanawal@u.nus.edu |  LinkedIn |  Google Scholar

EDUCATION

National University of Singapore (NUS)

PhD, Department of Electrical and Computer Engineering

Jan 2025 – Jan 2029

CGPA: 4.5/5.0

Birla Institute of Technology and Science, Pilani

Bachelor of Engineering, Electronics and Electrical Engineering

Aug 2018 – June 2022

CGPA: 8.39/10

RESEARCH PUBLICATIONS

Relevance as a Vulnerability: How Web Retrieval Degrades Safety Alignment in LLM Agents [PDF]

Aditya Nawal, Manit Baser, Mohan Gurusamy

arXiv preprint arXiv:2605.29224

May 2026

Built AgentREVEAL to diagnose retrieval-induced safety degradation in LLM agents; identified relevance-triggered failure modes including the Safe Source Paradox, and introduced HarmURLBench with 1,405 real-world URLs and 320 harmful behaviors.

Revolutionizing visuals: The role of generative AI in modern image generation [PDF]

Gaurang Bansal, Aditya Nawal, Vinay Chamola, Norbert Herencsar

ACM Transactions on Multimedia Computing, Communications and Applications

Nov 2024

Surveyed modern generative image-generation systems and analyzed their role in multimedia content creation, including quality, diversity, interpretability, computational efficiency, and practical deployment considerations.

An IoT-based incremental learning framework with a novel feature selection algorithm for intelligent diagnosis in smart healthcare [PDF]

Siva Sai, Kartikey Singh Bhandari, Aditya Nawal, Vinay Chamola, Biplab Sikdar

IEEE Transactions on Machine Learning in Communications and Networking

Mar 2024

Developed an IoT-based incremental learning framework for smart healthcare diagnosis, improving anomaly detection and classification accuracy while reducing runtime versus batch-learning baselines.

EXPERIENCE

National University of Singapore (NUS)

PhD Researcher, Department of Electrical and Computer Engineering

Jan 2025 – Present

- Researching safety vulnerabilities in **retrieval-augmented LLM agents** and developing robust AI alignment methodologies.
- Developed **AgentREVEAL** and **HarmURLBench** (1,405 URLs, 320 behaviors) to analyze how retrieved context degrades agentic safety.
- Investigated multi-agent collusion, introducing the **Trust Laundering** attack vector to demonstrate how adversarial consensus bypasses inline filters.
- Designed and evaluated multilingual agent baselines to analyze cross-lingual safety alignment and transfer in collaborative workflows.
- Conducted **mechanistic interpretability** on LLM agents to inspect internal representations and belief shifts during adversarial persuasion.

Microsoft

Software Development Engineer

July 2022 – December 2024

- Engineered **Azure OpenAI-assisted reasoning workflows** to automate network failure diagnosis, enhancing consistency and resolution speed.
- Developed production tooling using **C#** and **React** to integrate LLM capabilities into internal cloud infrastructure services.
- Automated deployment workflows and onboarding pipelines for network operations utilizing infrastructure-as-code automation.

Microsoft

Software Development Intern

Jan 2022 – June 2022

- Designed and implemented a centralized service marketplace platform to streamline internal tool discovery and cloud service usage.
- Consolidated fragmented software catalog services to improve developer experience and coordination across provider teams.

Akamai Technologies

May 2021 – July 2021

Software Development Intern

- Developed automated security tooling to test DNS query routing and validate packet handling integrity across security boundaries.
- Configured simulated environments utilizing the **TOR network and Onion routing protocols** for anonymity assessment.
- Engineered network packet captures (pcaps) replicating threat scenarios in the **MITRE ATT&CK framework** to test defense signatures.

PROJECTS

Copilot Skill For Network Troubleshooting and Remediation

Azure OpenAI

- Developed an intelligent **Copilot Skill** using **C#** and **Semantic Kernel** for autonomous network fault diagnosis via natural language.
- Built integration pipelines using Azure Functions to transform diagnostic telemetry into structured, actionable troubleshooting plans.

TECHNICAL SKILLS

Core Languages & Frameworks: Python, C++, PyTorch, JAX, TensorFlow, Git, LaTeX, C#, React, SQL

Research & Expertise: LLM Alignment (RLHF, DPO), AI Safety & Red Teaming, Adversarial Machine Learning, Mechanistic Interpretability, Multi-Agent Systems, RAG Security, Multilingual NLP

Agentic & LLM Frameworks: AutoGen, Semantic Kernel, Swarm, Magentic-One, LangChain, Hugging Face, OpenAI API

HONORS AND SERVICE

Leadership and Service: Led NSS BITS Pilani fundraising initiatives for underprivileged education and coordinated Microsoft IDC giving programs supporting child welfare, girl-child education, and disability-focused organizations.