ALLAN SURESH

AI Alignment Researcher | Software Engineer | Full Stack Developer

allansuresh9@gmail.com | +91 904-808-0008 | Pune, India | www.allansuresh.com
allan-suresh-824248147 | allansuresh

PROJECTS

Cybench Elicitation Ablation

Independent Research Project

- **Ran a controlled elicitation ablation** on UK AISI's Cybench cybersecurity benchmark (9 tasks, 6 categories) using Inspect AI, varying prompt guidance and agent tool access across two Claude models to separate genuine capability limits from harness artifacts
- **Found finished attempts were correct 100% of the time** (54/54 across 112 samples), showing apparent capability failure was mostly token-budget exhaustion, not incorrect reasoning – reported with an explicit confidence interval rather than an unqualified 100%
- **Identified a three-way task taxonomy** (reliably solved, reliably budget-exhausted, budget-critical) and confirmed the ambiguous middle category with a targeted 5x follow-up run
- **Diagnosed and fixed live infrastructure failures** (an under-provisioned token limit, mid-run API credit exhaustion) by reading raw eval logs rather than assuming results were valid

Sandbagging Detection & Elicitation Methodology Study

Independent Research Project

- **Extended a 2026 published finding** on sandbagging-induced "position attractors" (Cacioli et al.) from small open models to a frontier closed model (Claude Sonnet 4.5) using Inspect AI, testing a reasoning-permission axis the original authors explicitly flagged as unexplored
- **Found the effect replicates at frontier scale** only under forced-instant-answer elicitation, and identified that honest responses show a distinct, milder version of the same bias – meaning position-skew alone is insufficient evidence of sandbagging without a matched honest baseline
- **Caught and corrected two methodological flaws**: an analysis bug conflating correct and incorrect answers in an early result, and a same-family confound in a weaker-model-imitation control, resolved by re-testing against a cross-family reference model
- **Built reusable Inspect AI evaluation infrastructure**: custom scorers, a statistical analysis pipeline (chi-squared contingency testing), and a weaker-model-imitation control replicating Taylor et al.'s "Auditing Games for Sandbagging" methodology

Psychology in LLMs Research - Bystander Effect

AISI Fellowship, GeorgiaTech

In Progress

- **Leading multi-agent LLM research team** investigating the bystander effect across major LLM models with a systematic behavioral evaluation framework
- **Developed comprehensive scenario generation pipeline** using Distilabel and Claude to create 1000+ ethically-charged scenarios varying by group size (2-500), clarity level, and authority presence

EXPERIENCE

Research Collaborator

GCRI

Aug 2021 – Feb 2022 | Remote

- Researched AI Safety transfer techniques and cross-domain safety architectures
- Helped organize AI Safety conference connecting researchers with industry practitioners

Volunteer Content Writer

AISafety.info

Aug 2024 – Dec 2024 | Remote

- Authored article analyzing AI progress and safety considerations
- Participated in weekly editorial reviews to improve content quality

Java Developer

Citibank

Jul 2023 – Present | Pune, India

- Designed REST APIs for Alternative Investments platform processing high-volume transactions with 99.9% uptime
- Configured Jenkins CI/CD pipelines for multi-environment deployments with automated testing and rollback procedures
- Collaborated with API gateway team to analyze performance and implement systematic evaluation methodologies

iOS Developer

Citibank

Jul 2022 – Jun 2023 | Pune, India

- Implemented investment portfolio screens for APAC mobile banking applications
- Enhanced application stability through systematic debugging and cross-functional bug resolution

WORKSHOPS

AI Safety Workshop

AISI, Georgia Tech

2025

- **Built custom evaluation infrastructure** using Inspect AI framework with specialized scorers measuring intervention rates, context sensitivity, and behavioral categorization across multiple choice responses
- **Designed rigorous experimental methodology** with fixed seeds, controlled prompt formats, and statistical analysis (ANOVA, chi-squared) to test group-size effects and model capacity differences
- **Established novel research domain** bridging social psychology and AI alignment, investigating diffusion of responsibility and moral intervention patterns in multi-agent systems

ACHIEVEMENTS

GCRI Research Fellowship

Global Catastrophic Risk Institute

2021

Selected for a research fellowship at GCRI, following a 6-month research collaboration with Seth Baum(Co-Founder and Executive Director) contributing to AI risk assessment and global catastrophic risk mitigation.

FLI Junior AI Researcher

Future of Life Institute

2022

Engaged in AI existential risk discussions and contributed to FLI's initiatives on AI safety.

EDUCATION

B.Tech in Mechanical Engineering

National Institute of Technology, Karnataka

July 2018 – April 2022

- **CGPA**: 7.61
- **Minor**: Electronics and Communication Engineering

- Completed a six-week AI Safety workshop organized by the AI Safety Initiative (AISI) at Georgia Tech covering foundational and advanced topics in AI alignment, multi-agent safety, and governance.
- Collaborated on practical exercises in risk assessment, robust alignment methods, and interpretability frameworks.

AI Safety Fundamentals Course

BlueDot Impact

2024

- Completed an in-depth six-week AI safety fundamentals course covering technical AI alignment and governance.
- Worked on assignments and case studies on technical studies in AI alignment and risk mitigation.

AI Safety Summer Workshop

SERI MATS

2023

- Completed John Wentworth's research-focused AI alignment remote workshop on Agent Foundation.
- Replicated and studied Hessian and Basin volume model by Vivek Hebbar.
- Engaged in daily discussions and group presentations on different alignment topics each session.

SKILLS

Python | Java | Docker | Git | SQL | Flask