Akash Kundu

Kolkata, West Bengal, India akashkundu2xx4@gmail.com +916289481664 in/akash-kundu-a334b1250

EDUCATION

Bachelor of Technology, Computer Science and Engineering

Heritage Institute of Technology • Kolkata • Oct 2022-Jun 2026

EXPERIENCE

Data Scientist

Humane Intelligence **January 2025 - Present, Remote**

- Upcoming Data Scientist at Humane Intelligence, contributing to the Red Teaming Evaluations team by analyzing post-event red-teaming data and preparing comprehensive reports.

Mentee

Berkeley AI Safety Initiative for Students **October 2024 -Present, Berkeley**

- Working as the sole mentee under Rohan Subramani to conceptualize a framework around Goals in Foundational Model Agents.
- Aiming to publish multiple blog posts on LessWrong and an arXiv preprint by early 2025.
- The project was eligible for **$2000** as compute reimbursements, though funding wasn't required.

Research Intern

AI Institute of South Carolina **September 2024 - Present, South Carolina**

- Collaborating with Professor Amitava Das on a research paper focused on Constitutional AI and prompt biases.
- Selected as one of **13** interns from a pool of **666** applicants.

Research Fellow

Apart Research **January 2024 - Present, San Francisco, Bay Area**

- Co-authoring a paper researching on cross-lingual capabilities of Safety Evaluations Benchmark.
- Co-authored a paper evaluating Dark Patterns in State-of-the-Art LLMs, including the development of a **600+** dataset to benchmark dark patterns across 6 distinct dark patterns, with an Arxiv preprint forthcoming and plans to scale the dataset.
- Both papers are currently under review at **ICLR**, **AAAI** and **ACL**.

Research Intern

Lionheart Ventures **June 2024 - September 2024, San Francisco Bay Area**

- Developed a **comprehensive framework** to evaluate the impact of **AI-induced systemic risks** across **10+** portfolio companies.
- Leveraged **Claude Sonnet** and **GPT-4o-mini** to generate over **5,000** scenarios, analyzing potential systemic risks from AI that could lead to existential threats.
- Performed **Cluster Analysis** and **Monte Carlo Simulations** to assign risk weights to **600+** distinct AI-related threats.
- Drew on methodologies similar to the MIT's AI Risk Initiative (https://airisk.mit.edu/), with a distinct focus on systemic risks leading to existential outcomes.
- The framework has become a core component of internal due diligence, enabling investing stakeholders to effectively evaluate and prioritize funding for high-impact organizations aligned with risk mitigation goals.

RESEARCH

Towards Smart Nation Rankings: A Data-Centric Approach Using Global Development Indicators

tinyurl.com/32rv4ytw • October 2024 -November 2024

- Co-authored a paper to measure the "smartness" of nations utilizing **1400+** features.
- Ranked **157** nations through this proposed taxonomy, establishing one of the most comprehensive and inclusive assessments of national "smartness" to date.
- Currently under review for ICAA 2025.

An Approach to Detect and Classify Potentially Suspicious Activity from Real-Time Log Data using Anomaly Detection Methods

IEEE INOCON (International Conference for Innovation in Technology) • ieeexplore.ieee.org/document/10511679

- November 2023 - December 2023
- Authored a paper highlighting an approach to detect and classify suspicious activity from real-time log data using unsupervised learning methods.

VOLUNTEERING

AI/ML Lead

Heritage institute of Technology • Google Developers Student Club Heritage Institute of Technology • July 2023 - June 2024

- Led hands-on ML coding sessions, engaging over 150 active participants.
- Launched our institute's first ML Hackathon, successfully hosting **300+** participants.
- Returned as the **Tech Lead** for 2024-25, overseeing the club's tech domain, with a team of 14 domain mentors across web development, Android, and ML. Currently supporting a community of over **1800** active members.

Volunteer ML Engineer

San Francisco, Bay Area • Omdena • February 2023 - October 2023

- Volunteered on various ML projects in Omdena.
- Developed a deepfake detection system for women safety in Germany.
- Estimated out-of-pocket lung cancer costs for patients in the US with a confidence interval of 10%.
- Created a matching algorithm for hostel roommates based on personality traits in Egypt.
- Fine-tuned Vicuna 13B for a mental health chatbot supporting English and Swahili, aimed at assisting people in Tanzania.

CERTIFICATION

AI Safety Fundamentals - Alignment Course

BlueDot Impact • 2024

- Participated in weekly discussions and learnt about Alignment Research and Technical AI Safety
- Developed a project to highlight the deceptive capabilities of LLMs. Showed that SoTA LLMs are incapable of collaborative deception. (https://www.apartresearch.com/project/werewolf-benchmark)

SKILLS

Languages: Python, C
Libraries: Pandas, Numpy, Matplotlib, Seaborn, Scikit-Learn, TensorFlow, Pytorch, BeautifulSoup, OpenCV, Albumentations, Keras, YOLO, NLTK, Huggingface transformers, Langchain
Data Visualization: PowerBI, Tableau
Deployment: Streamlit
Miscellaneous: SQL, Excel, PowerPoint

ACHIEVEMENTS

Accepted in CaMLAB V4, an ARENA equivalent ML BootCamp organized by Cambridge AI Safety Hub
Received CoAuthor Status for MMTEB (massive multilingual text embedding benchmark) based on my open-source contributions to their repository. MMTEB is currently under review for **ICLR.**
Shortlisted for Round 3 of the Atlas Fellowship India, 2022.
Winner of MLTiverse, Manipal University Jaipur out of **600+** participants.
Ranked **1st** out of **3000** participants on Kaggle in Kharagpur Data Science Hackathon by IIT Kharagpur
Winner of LLM Evaluations Hackathon (November '23) and AI Security Evaluations Hackathon (May '24) organized by Apart Research.
1st Runners up at ClimateConnect Hackathon, IEEE, Jadavpur University.
Won the Spectral Syntax Bounty worth **$2000** in the Web3 x AI Hackathon by Encode Club in July 2024 for building Decentralized AI Agents
Secured 1st position in the Concordia Hackathon hosted by Apart Research and Co-operative AI Foundation in September 2024.
Received **$1100** from Google Deepmind as Researcher credits for the Neurips Concordia Competition, 2024.