Anjila Budathoki

✉ anzilabudathoki@gmail.com, ☎ 404-203-7842
https://anjilab.github.io/ linkedin.com/in/anjilabudathoki/

Interest Area: AI Safety, Jailbreaking, Blindspots in AI safety training, AI in decision-making, HCI, AI in social science, Applied AI systems, Knowledge Distillation

About

I am a third-year Ph.D. student at Georgia State University, with primary research interests at the intersection of AI safety and human-centered AI. My work focuses on understanding and enhancing the trustworthiness and safety of AI systems, including the study of adversarial attacks and defenses, as well as how humans perceive, interact with, and are affected by AI systems. I aim to investigate how AI systems, designed by humans and for humans can collaborate with people in safe, reliable, and socially responsible ways. More broadly, I am inclined toward interdisciplinary research that integrates psychology, social science, linguistics, and computer science to create meaningful societal impact.

Education

Georgia State University August 2023-Present
PhD in Computer Science
Overall Grade: **3.99**

Advanced College of Engineering & Management, Lalitpur, Nepal 2015-2019
Bachelor's Degree in Computer Engineering
Graduation: February 2020
Letter Grading: A^+

Publications

Peer Reviewed

- Jayden Fassett, **Anjila Budathoki**, Jack Morris, Qin Hu, and Yi Ding. "Unobtrusive Universal Acoustic Adversarial Attacks on Speech Foundation Models in the Wild." In Proceedings of the 27th International Conference on Multimodal Interaction, pp. 424-433. 2025.
- Deniz Marti, **Anjila Budathoki**, Yi Ding, Gale Lucas, and David Nelson, "How does acknowledging users' preferences impact AI's ability to make conflicting recommendations?" In the International Journal of Human–Computer Interaction, 2024. - *International Journal of Human–Computer Interaction 2024*

Work in Progress/ Under Review

- **Understanding the Role of Prompt Template in Knowledge Distillation for Safety Alignment**
  Submitted to ACL Short Paper 2026
  - Investigating the role of prompt templates in distillation,
  - Exploring task utility vs safety impact of prompt template in distillation
- **Persuasive behavior of Personality Induced LLM arguments**
  - Investigating how specific personalities (Big Five - OCEAN) in large language models (LLMs) influence argument generation for polarized and non-polarized topics,
  - Impact on participants' opinion change after exposure to personality specific arguments.,
  - Analysing perception of influence, persuasiveness, and source of generated arguments.
- **Exploring the safety alignment of efficient techniques in language models**
  - Investigating the impact of safety alignment in finetuning and distillation methods.
  - Exploring task utility vs safety of models tradeoff in efficient methods of language models.
  - Focusing on safe distillation of alignment techniques from larger aligned models.

Industry Experience

Summer Intern 2024 at Toyota Infotech Labs

- Investigation of utilizing personalization in electric vehicles,
- Analysis of speed prediction based on recommendation-based algorithm.

YoungInnovations Pvt Ltd. April 2019-July 2023
Mid Software Engineer

- Utilized the feature of the Places library in the Maps Javascript API. Mainly, focused on Autocomplete features,
- Designed and refactored the UI, especially the Chart and Table component's logical part.For charts, Apexcharts were used,
- Database design using tools like db diagram.io,
- Code review of colleague,
- Team lead for management of JavaScript projects,
- Taking interview, guidance to interns.

Projects

- **Breast Cancer Prediction Using Neural Networks**: Developed and evaluated a convolutional neural network (CNN) using PyTorch for classifying histopathological breast cancer images into benign and malignant categories (Undergraduate project).
- **Nepal Project Bank Management Information System**: Developed an information system for the Government of Nepal using Node.js of backend services, Express for REST API and React for the frontend, enabling digital tracking and management of national project status.
- **Nepal Census Data Analysis System**: Developed a data analysis and visualization system using Node.js, Vue.js, PostgreSQL, and Prisma to process, visualize, and share housing and population data from the Nepal Census for three municipalities.
- **Sombar**: Designed and implemented an employee management information system using Node.js, React, and PostgreSQL, supporting attendance tracking, leave management, and team-level dashboards.
- **Skill Lab – Career Service Center**: Developed a job-matching management platform using Next.js, TypeScript, Laravel, and Tailwind CSS to facilitate connections between skilled students and hiring organizations.
- **Avo Portal**: Implemented a delivery service portal using Nuxt.js with Google Maps integration to support location-based logistics and goods delivery.
- **E-Sifarish**: Built an information system for local government ward offices using Node.js, React, PouchDB, and Knex.js to digitize application workflows.
- **VoiceInn**: Assisted in a web-based service application using Vue, Python, and TypeScript; implemented backend using python, including a Windows Service application, and conducted performance optimization of jQuery widgets.

Skills

Programming Languages:	JavaScript, TypeScript, Python, C++,
Frameworks:	PyTorch, Express, Django, Flask, Tailwind CSS
Libraries:	Numpy, Pandas, Matplotlib, Seaborn, React, Next.js
Platforms:	Linux, Ubuntu, Docker,

Course Taken/Taking

- CSC 6850: Introduction to machine learning
- CSC 6851: Introduction to deep learning
- CSC 8850: Advanced Machine learning

- CSC 8851: Deep learning
- CSC 8370: Data security
- CSC 8980: Natural Language Processing
- CSC 8260: Advanced Image Processing
- CSC 8230: Secure and Private AI

Community Involvement

- Participated in Project Association for Computer and Electronics (PACE) club of ACEM Hackathon
- Participated in Hack Day 2021 YoungInnovations Pvt. Ltd
- NASCOIT Paper Presentation