Anjila Budathoki

anzilabudathoki@gmail.com, 404-203-7842, Atlanta,Georgia
github.com/anjilab, linkedin.com/in/anjilabudathoki/

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

About

I am a third-year Ph.D. student at Georgia State University, with a primary research focus on natural language processing and AI in decision-making. My primary research interests focus on understanding and enhancing the trustworthiness and safety of AI systems. This includes studying adversarial attacks and defenses as well as exploring how humans interact with these systems. I aim to leverage machine learning techniques to tackle complex challenges and develop robust, reliable models. Furthermore, I am inclined to interdisciplinary research integrating 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: **4.07**

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- October 2019
Software Developer Intern

- Designed and developed a REST API.

YoungInnovations Pvt Ltd. November 2019- February 2020
Jr. Software Developer/Engineer

- SQL And NoSQL,
- Code refactor

YoungInnovations Pvt Ltd. March 2020- 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

1. **Breast Cancer Prediction Using Neural Network**
   Undergraduate project
   - Implementation of deep learning based algorithm i.e Convolutional Neural Network in breast cancer prediction,
   - To identify the breast cancer condition whether it's benign or malignant based on the histopathological images.
2. **Nepal Project Bank Management Information System**
   Node.js, React, Socket.io
   - Information system for Nepal government about the projects
   - Designed and developed a REST API.
3. **Census Data Visualization**
   Node.js, Vue.js, PostgreSQL, Prisma
   - Development of system to analyze the housing and population data of Nepal Census and visualize and share it for 3 municipalities.
4. **Voiceinn**
   Vue, Python, Typescript
   - Implemented a web service application embedded in a toolbox using technologies such as C#, JavaScript, SQL, ASP .NET and exceeded clients' expectations by optimizing jQuery widgets.
   - Designed and developed a Windows Service application and obtained positive feedback from clients by effectively communicating the client needs and executing tasks efficiently.
   - Obtained a full-time offer from the Statistics Information System Division (SISD) executive team by demonstrating strong self-learning skills and work ethic.
5. **Sombar**
   Node.js, React, PostgreSQL

- In house project for employee management information system regarding their attendance, leaves, respective dashboard for team leads.

6. **Skill Lab - Career Service Center**
   Next.js, Typescript, Laravel, Tailwind CSS
   - Management system that intends to build a relationship among organizations which provides jobs and students that are skilled to take those jobs.
7. **Avo portal**
   Nuxt.js, Google maps
   - Portal site for delivery of goods
8. **E-sifarish**
   Node.js, React, PouchDB, knex.js
   - Management system for local government ward offices to handle applications/sifarish.

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