Akshita Jha
https://akshitajha.github.io
Email : akshitajha@vt.edu
LinkedIn: linkedin.com/in/akshitajha/

Education

- **Virginia Tech** — Washington DC-Baltimore Area
  Ph.D., Dept. of Computer Science; GPA: 4.00 — *Aug 2019 – Dec 2024*
  - **Thesis**: Adversarial Risks and Stereotype Mitigation at Scale in Generative Models
  - **Committee**: Dr. Chandan K. Reddy (chair), Dr. Lifu Huang, Dr. Xuan Wang, Dr. Su Lin Blodgett, Dr. Vinodkumar Prabhakaran
- **IIIT-Hyderabad** — Hyderabad, India
  MS by Research, Natural Language Processing
- **IIIT-Hyderabad** — Hyderabad, India
  Bachelor of Technology (Hons.), Computer Science

Research and Work Experience

- **Apple**
  Research Scientist (Contract), Apple Services Engineering — *Jan 2025 - Present*
  - **Scaling Safety of Generative Models**: Scaling Responsible AI approaches to ensure safe releases of the rapidly evolving AI models.
- **Google Research** — Mountain View, CA
  Research Intern, Responsible AI — *May 2023 - August 2023*
  - **Safety of Text-to-Image Models**: Led a global scale analysis of visual stereotypes in Text-to-Image models, like Stable Diffusion. Developed a multi-faceted evaluation paradigm for detecting regional stereotypes in generated images of 135 identity groups worldwide. Assessed the extent to which stereotypes are reproduced in images.
- **Google Research** — Mountain View, CA
  Research Intern, Responsible AI — *August 2022 - January 2023*
  - **Safety of Large Language Models (Foundation Models)**: Developed a broad-coverage stereotype benchmark by leveraging generative capabilities of PaLM, and GPT-3 using few-shot prompting. Evaluated several large language models (LLMs) for the presence of regional stereotypes. Ensured robust validation by engaging a globally diverse rater pool to assess their prevalence in society – both within and outside their region of origin.
- **AI Lab, InterDigital** — Palo Alto, CA
  Research Intern — *May 2020 - August 2020*
  - **Interpretable Long Document Matching**: Developed a contrastive learning framework to compute (dis)similarity within and across different chunks and sections of long documents in an interpretable manner using contrastive learning along with BERT embeddings and custom position embeddings.
- **Virginia Tech** — Arlington, VA
  Research Assistant, Advisor: Chandan K. Reddy — *August 2019 - Present*
  - **Inherent Biases vs. Task-Specific Limitations in Generative Models**: Conducted a large-scale evaluation to identify whether stereotypical disparities in model outputs stem from unfair biases against identity groups or the model's (in)ability to handle specific downstream tasks – providing crucial insights for mitigation.
  - **Robustness of Code Programming Language Models**: Designed 'CodeAttack', a black-box attack model tailored for assessing the robustness of programming language models at scale. Leveraged code structure to generate effective, efficient, and imperceptible adversarial code samples that can be used for red-teaming. Successfully exposed vulnerabilities of various models across tasks like code translation, code repair, and code summarization.
  - **Challenges in Transformer-based Models for Long-Form Document Matching**: Evaluated the effectiveness of simple neural models and simple embeddings over transformer-based models on the task of document matching. Empirically demonstrated that simple models are at par with the more complex BERT-based models while taking significantly less training time, energy, and memory.
  - **Fair Representation Learning**: Built a model for learning fair disentangled representations while ensuring the utility of the learned representation for downstream tasks.
- **Google Summer of Code, Debian** — Remote, India
  Software Engineering Intern — *May 2015 - August 2015*

- **The Linux Foundation, OpenDaylight** — Remote, India
  Software Engineering Intern, Deb Maintainer — *May 2016 - April 2017*
- **Intuit** — Bangalore, India
  Software Engineer I — *August 2017 - May 2018*

Selected Publications

For the most recent list, please refer to my Google Scholar.

- **Akshita Jha**, Sanchit Kabra, Chandan K. Reddy. *"Biased or Flawed? Mitigating Stereotypes in Generative Language Models by Addressing Task-Specific Flaws"* Under Review
- Sanchit Kabra, **Akshita Jha**, Chandan K. Reddy. *"Reasoning Towards Fairness: Mitigating Bias in Language Models through Reasoning-Guided Fine-Tuning"* Under Review
- **Akshita Jha**, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Chandan K. Reddy, Sunipa Dev. *"ViSAGE: A Global-Scale Analysis of Visual Stereotypes in Text-to-Image Generation"* ACL 2024
- **Akshita Jha**, Aida Davani, Shachi Dave, Vinodkumar Prabhakaran, Sunipa Dev. *"SeeGULL: A Stereotype Benchmark with Broad Geo-Cultural Coverage Leveraging Generative Models"*, ACL 2023
- Sunipa Dev, **Akshita Jha**, Jaya Goyal, Dinesh Tewari, Shachi Dave, Vinodkumar Prabhakaran. *"Building Stereotype Repositories with Complementary Approaches for Scale and Depth"*, Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), EACL 2023
- **Akshita Jha** and Chandan K. Reddy. *"CodeAttack: Code-based Adversarial Attacks for Pre-Trained Programming Language Models"*. AAAI 2023. **[Spotlight Presentation]**
- **Akshita Jha**, Adithya Samavedhi, Vineeth Mohan, and Chandan K. Reddy. *"Transformer based Models for Long Document Comparison: Challenges and Empirical Analysis"*, EACL (Findings), 2023
- **Akshita Jha**, Vineeth Mohan, Jaideep Chandrashekar, Adithya Samavedhi, and Chandan K. Reddy. *"Supervised Contrastive Learning for Interpretable Long-Form Document Matching"*. ACM Transactions on KDD, May 2022
- **Akshita Jha** and Radhika Mamidi. *"When does a Compliment become Sexist? Analysis and Classification of Ambivalent Sexism using Twitter Data"*. Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science, ACL 2017 **[Spotlight Presentation]**

Awards

- **Kafura Graduate Fellowship 2024**: Awarded to CS PhD Students for overall excellence and outstanding research
- **Travel Awards**: AAAI 2023, ACL 2023, EACL 2023
- **CS Research Mentorship Program Scholar**: Google 2021
- **Grace Hopper Scholarship**: Grace Hopper Scholar for the year 2020
- **Member of Phi Kappa Phi Honor Society**: Awarded to top 1% for academic excellence

Service

- Reviewer for several machine learning (ML), artificial intelligence (AI), and Natural Language Processing (NLP) conferences like ACL, AAAI, EACL, NAACL, KDD, CVPR, etc., and journals like TKDD, and IEEE.

Programming Skills

- **Languages and Technologies**: Python, PyTorch, TensorFlow, Keras

Research Interests

- Large Language Models (LLMs), Safety, Red-teaming, Alignment, Generative AI, Fairness and Robustness of Foundation Models, Natural Language Processing (NLP), Dialog Models