Veda Duddu

LinkedIn | veda.duddu@gmail.com

EDUCATION

University of Illinois Urbana- Champaign **USA**
Incoming PhD in Computer Science *Aug 2026*

University of Illinois Urbana- Champaign **USA**
MSCS *Aug 2024 - May 2026**

- **GPA:** 3.95/4.00
- **Focus:** AI Safety, Human-AI Interaction, Sociotechnical AI Systems. **Advised by Koustuv Saha, onCare Lab**

Ashoka University **India**
Bachelor of Science (Honors) - Mathematics and Computer Science *Aug 2019 - May 2023*

- **GPA:** 3.76/4.00
- **Focus:** Human-AI Interaction, Sociotechnical AI Systems, Social Media Analysis

PUBLICATIONS

- **Duddu, V.,** Parekh, J. R., Mao, A., Min, H., Xiao, Z., Das Swain, V., Saha, K. **(2026)**. Not My Truce: Personality Differences in AI-Mediated Workplace Negotiation. arXiv:2604.00464. Under review**.** **[**Preprint Link**]**
- **Duddu, V.*,** Pal, O.*, Goyal, A., Goel, D., Saha, K. **(2026).** Do We Know What They Know We Know? Calibrating Student Trust in AI and Human Responses Through Mutual Theory of Mind. arXiv:2601.16960**.** *CHI 2026 EA.* (*equal contribution) [Preprint Link]
- **Duddu, V.,** Parekh, J. R., Mao, A., Min, H., Xiao, Z., Das Swain, V., Saha, K. **(2025).** Does AI Coaching Prepare us for Workplace Negotiations? arXiv preprint arXiv:2509.22545. *Under review.* [Preprint Link]
- Liu, Z., Rabbani, P**., Duddu, V.,** Fan, K., Lee, M., Huang, Y. **(2025).** The Social Gaze of LLMs: A Literature Review of Multimodal Approaches to Human Behavior Understanding. arXiv preprint arXiv:2510.23947. *Under review.* [Preprint Link]

RESEARCH EXPERIENCE

onCare Lab - UIUC **USA**
Graduate Research Assistant *Nov 2024 - Present*

Trucey AI Negotiation Coach Study (N=267)

- Investigated scalable oversight failure mode through 267-person experiment; conversational AI reduced fear but degraded human judgment capacity, query-response interaction imposed cognitive load under stress, undermining integrative thinking
- Identified critical alignment challenge, as AI systems become more naturalistic (conversational vs static), they paradoxically reduce human supervisory capacity (6.5× weaker empowerment)
- Demonstrated context-dependent safety failures: cognitive profiles predicted when AI guidance helped vs harmed; structured support conflicted with flexible reasoning needs in high-stakes scenarios

AI-Assisted Education Study (N=8)

- Revealed trust-reliance disconnect in AI-assisted contexts, users pragmatically relied on AI despite accuracy concerns, demonstrating oversight limitations and strategic autonomy preservation

SALT Lab - UIUC **USA**
Research Assistant *Jan 2025 - Sep 2025*

Multimodal AI Safety Evaluation Study (N=176)

- Engineered LLM-assisted research pipeline (Gemini Pro, few-shot prompting) to systematically evaluate safety practices in 176 multimodal AI systems. Validated methodology achieved κ=0.717 agreement
- Identified critical safety gaps, 93% use modality-to-text conversion stripping social context (misinterpretation risk), 45% lack ethical discussion, evaluation over-relies on static benchmarks vs human-centered assessment

Algoverse - Winter 2024 Cohort **USA**
Research Assistant *Dec 2024 - Mar 2025*

RL Agent Misalignment Study (Minecraft Testbed)

- Conducted experiments on model organisms of misalignment in Minecraft RL environments; empirically documented alignment failures including mesa-optimizer objective resistance, reward hacking via underground tunneling, and instrumental convergence failures; co-authored LessWrong safety research article [Blog]

Social Network Research Group, Ashoka University x Mphasis Labs **India**
Undergraduate Research Assistant *May 2022 - May 2023*

- Investigated bias in the production and consumption of Indian media by analyzing tweets and articles under Professor Debayan Gupta, building understanding of how cultural context shapes subjective content and its interpretation at scale
- **Synthesised** a pipeline to produce over **100,000+ articles from 5 Media Houses in 30 hours** which will be contributed to produce a formal database. **Designed visualisations** to understand hashtag usage of 12 Media Houses

TEACHING EXPERIENCE

Siebel School of Computing and Data Science **USA**
Graduate Teaching Assistant - CS 173, Discrete Mathematics *Jan 2026 - Present*

- Facilitated class tutorials and graded 60+ student problems weekly, rotating coordination of full TA team examlet grading once per semester under Professor Carl Evans and Professor Margaret Fleck

Siebel School of Computing and Data Science **USA**
Graduate Teaching Assistant - CS 447, Natural Language Processing *Aug 2025 - Dec 2025*

- Teaching Assistant to **Professor Hao Peng, Fall 2024**
- Conducted office hours for 400+ students, providing conceptual guidance and debugging support for NLP projects

CS Department, Ashoka University **India**
Undergraduate Teaching Assistant - Introduction to Computer Programming *Jan - May 2022*

- **Assisted** and **facilitated** learning experiences for **Professor Subashis Banerjee** through office hours, construction and grading of all forms of assessments in Introduction to Computer Programming for **over 170 students**, evaluated as **4.41/5.00**
- **Awarded** Undergraduate Teaching Excellence Award by the CS Department out of **30+ Teaching Assistants** for helpful and dedicated instruction

INTERNSHIP EXPERIENCE

University of California Santa Cruz *Jul 2023 - Aug 2023*
Summer Research Intern

- Investigated nine causal inference frameworks for policy evaluation under Professor Zehang Richard Li, implementing methods including **Instrumental Variables** and **Difference-in-Differences** in R.
- Translated theoretical tools into **reproducible implementations**, building rigorous evaluation skills for studying causal effects across diverse populations.

Nichesolv **India**
Data Science Intern *Jun - August 2022*

- **Improved annotation consistency** for a tennis shot classification model by **labeling 7,000+ data snippets**, building firsthand understanding of how labeling decisions introduce subjectivity and inconsistency at scale.
- **Strengthened model reliability** by conducting error analysis across **10+ experiments on the LRCNN mode**l for predicting tennis players' strokes, identifying systematic failure patterns and quantifying model uncertainty across diverse stroke types

POSITIONS HELD

CS Department Ashoka University, Academic Advisory Board **India**
Computer Science Representative *Apr 2021 - Apr 2022*

- Facilitated logistical communication between 300 students and the CS Department, mentoring 35 students by designing their course trajectories and guiding their CS experiences.
- Advocated for inclusivity for interdisciplinary students by incorporating dedicated sections into the Department Handbook, recognized as 1 of 4 recipients of the Undergraduate Service Excellence Award out of 300+ students as the inaugural CS-Math interdisciplinary representative

Women in Computing Society, Ashoka University **India**
Head of Blog Team *Jun 2021 - May 2022*

- Led a team of 12 to advocate for women in STEM, editing and authoring articles showcasing accomplishments of women in computing, and designed the inaugural WiCS newsletter Bit a Bit from scratch

Events and Podcast Member *Sep 2020 - Jan 2022*

- Led a Workshop Weekend introducing **R** to non-technical majors, designing accessible curriculum with no prerequisite knowledge assumed, and helped ideate 3 additional workshop weekends across diverse technical topics.

PROJECTS

Coworking Illini - Study Buddy Matching App
CS465: User Interface Design *Aug 2024 – Dec 2024*

- **Led front-end development** and **UX validatio**n for a study buddy matching platform, conducting 3 rounds of usability testing using think-aloud protocols and task analysis across diverse user groups, achieving **80% user satisfaction**
- **Translated iterative user feedback** into interface refinements, building experience in human-centered design and behavioral observation methods

(Dis)Passion of Deduction
CS2109/ENG2350: Introduction to Digital Humanities *Feb 2021 - Mar 2021*

- **Analysed** over **9 books** using Natural Language Processing's **Named Entity Recognition** to understand and question the existence of the stereotype of detectives in 19th-century crime fiction being cold and rational individuals through **sentiment analysis** [Github Repo]

TECHNICAL SKILLS

- **Programming Languages**: Python (Advanced), JavaScript, R, C
- **ML/AI Frameworks**: PyTorch, TensorFlow, Keras, scikit-learn, Hugging Face Transformers
- **Research Methods:** Controlled Experiments, Statistical Analysis (regression, effect sizes, t-tests), Survey Design, Inter-Rater Reliability (IRR), Qualitative Coding, RL Experimentation