

Vamshi Krishna Bonagiri

[✉ vamshi.12.2003@gmail.com](mailto:vamshi.12.2003@gmail.com) [🌐 Homepage](#) [LinkedIn](#) [G Google Scholar](#)

Education

| | | |
|--------------|--|--------------|
| 2025–present | PhD in Machine Learning <i>Mohammed Bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE</i> | GPA: -/- |
| 2020–2025 | B.Tech. in Computer Science and M.S. by Research in Computational Linguistics, <i>International Institute of Information Technology, Hyderabad</i> | GPA: 9.01/10 |

Research Experience

| | |
|-----------|---|
| 2025-2025 | Center for Human-Compatible AI (CHAI), UC Berkeley , <i>Research Intern, USA</i> Worked at Stuart Russell's Lab under the supervision of Benjamin Plaut on evaluating and improving LLM Agent uncertainty quantification methods for complex multi-turn settings. |
| 2022-2024 | Microsoft Research , <i>Visiting Researcher, India</i> Collaborated with Monojit Choudhury on creating a benchmark on code-mixed acceptability and with Tanuja Ganu on developing cross-modal adversarial attacks. |
| 2021-2025 | Precog, IIIT Hyderabad , <i>Research Assistant/Undergraduate Researcher, India</i> Worked on NLP and AI Safety Research under the supervision of Ponnurangam Kumaraguru , Here's a list of published research . |
| 2023-2024 | KAI² Lab, University of Maryland , <i>Undergraduate Research Assistant, USA</i> Worked on LLM Evaluations for reliability and trustworthiness, led by Manas Gaur . |
| 2021-2022 | Tokyo Institute of Technology , <i>Research Intern, Japan</i> Analyzed a corpus of 87,000 Reddit comments related to Deepfakes with Sasahara . |

Industry Experience

| | |
|------|--|
| 2024 | Sprinklr , <i>Machine Learning Product Engineering Intern, Gurgaon, India</i> Developed a custom LLM-based Multi-Agent framework (similar to Autogen) with 30% higher accuracy, 85% latency reduction, and 80% cost reduction compared to off-the-shelf solutions, accelerating sales analysis by 28x. |
| 2022 | Observe.ai , <i>Machine Learning Research Intern, Bangalore, India</i> Developed a low-resource multilingual text classification architecture with a 75+ F1 score, enabling product adaptation from English to Spanish. |
| 2022 | Smart City Living Labs , <i>Software Engineer Intern, Hyderabad, India</i> Developed and deployed an immersive 3D dashboard for real-time monitoring of 300+ IoT devices across the IIIT Hyderabad campus. |

Teaching & Mentoring

| | |
|-----------|--|
| 2024-2025 | IIIT Hyderabad & NPTEL , <i>Teaching Assistant</i> (1) Using AI Tools for Research Workshop at IIT Hyderabad (2025) (2) Responsible & Safe AI NPTEL course (2024) (3) Responsible and Safe AI Systems course, supported by Open Philanthropy grant (2024) (4) Introduction to NLP (2023) |
|-----------|--|

| | |
|------|--|
| 2024 | ACM India Summer School on Responsible & Safe AI , <i>Teaching Assistant</i> , IIT Madras |
| 2024 | Bluedot Impact , <i>AI Safety Fundamentals Facilitator</i> , United Kingdom Led an international cohort in the AI Safety fundamentals course . |

Grants & Awards

| | |
|-----------|---|
| 2023-2025 | Research Distinctions: OpenAI Research Access Grant: 5000\$, Google DeepMind Symposium 2025 (top 200 ML scholars in India), IndoML 2024 (top 20 ML graduate scholars), Research Award, IIIT Hyd (Spring 2024), Open Philanthropy AI Safety Course Grant, AI Alignment Workshop Grant by FAR.AI , Black Box NLP Grant , Global Challenges Project |
| 2020-2024 | Academic Awards: Dean's List 2 (Spring 2024, 2022), Dean's List 1 (Monsoon 2021), Merit List (Spring 2021, Monsoon 2020) |
| 2017-2021 | Competition Awards: Megathon 2021 Winner (COVID mask detector), NASA AMES Space Settlement Contest 2017 (Top 3 internationally) |

Leadership & Service

| | |
|-----------|---|
| 2021-2023 | Overall Co-ordinator , Open Source Developers Group, IIIT Hyderabad |
| 2020-2022 | Tech Team Member , Entrepreneurship Cell, IIIT Hyderabad |
| 2024-2025 | Volunteer , Mental Health Support Group |
| 2020-2025 | Freelancer - Projects in blog writing, book reviews, poetry generation, web development, and Machine Learning tutoring |

Papers and Publications

| | |
|------|---|
| 2025 | <i>Check Yourself Before You Wreck Yourself: Selectively Quitting Improves LLM Agent Safety</i> Bonagiri, V.K. , Kumaraguru, P., Nguyen, K., Plaut, B. Regulatable ML and Reliable ML Workshops , NeurIPS 2025 |
| 2025 | <i>Dark Side of the Tune: Investigating the maladaptive outcomes of excessive music consumption in the age of unlimited music access</i> Bonagiri, V.K. , Alluri, V. 18th International Conference on Music Perception and Cognition (ICMPC) 2025 |
| 2024 | <i>Evaluating Moral Consistency in Large Language Models</i> Bonagiri, V.K. , Vennam, S., Govil, P., Kumaraguru, P., and Garg, M. LREC-COLING 20-25 May, 2024, Torino, Italy |
| 2025 | <i>If Pigs Could Fly... Can LLMs Logically Reason Through Counterfactuals?</i> Balappanawar, I.B.*, Bonagiri, V.K. *, Joishy, A.R.* , Gaur, M., Thirunarayan, K., Kumaraguru, P. Under Review 2025 |
| 2025 | <i>From Human Judgements to Predictive Models: Unravelling Acceptability in Code-Mixed Sentences</i> Kodali, P., Goel, A., Bonagiri, V.K. , Choudhury, M., Shrivastava, M., Kumaraguru, P. ACM TALLIP (Journal) 2025 |
| 2025 | |

COBIAS: Contextual Reliability in Bias Assessment
Govil, P., **Bonagiri, V.K.**, Garg, M., and Kumaraguru, P
ACM Web Science 2025

- 2024 *K-PERM: Personalized Response Generation Using Dynamic Knowledge Retrieval and Persona-Adaptive Queries*
Raj, K., Roy, K., **Bonagiri, V.K.**, Thirunarayanan, K
AAAI-MAKE 2024
- 2023 *Representation Learning for Identifying Depression Causes in Social Media*
Govil, P., **Bonagiri, V.K.**, Garg, M., and Kumaraguru, P
Proceedings of KDD Kil 2023. Long Beach, California, USA August 6, 2023
- 2023 *Towards Effective Paraphrasing for Information Disguise. In European Conference on Information Retrieval (pp. 331-340)*
Agarwal, A., Gupta, S., **Bonagiri, V.K.**, Gaur, M., Reagle, J., Kumaraguru, P
ECIR March, 2023
- 2022 *Are deepfakes concerning? analyzing conversations of deepfakes on Reddit and exploring societal implications. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1-19)*
Gamage, D., Ghasiya, P., **Bonagiri, V.K.**, Whiting, M. E., Sasahara, K
CHI April, 2022