

Vamshi Krishna Bonagiri

✉ vamshi.12.2003@gmail.com  [Personal Homepage](#)  [LinkedIn](#)  [Google Scholar](#)

Education

2020–2025 **B.Tech. in Computer Science and M.S. by Research in Computational Linguistics,**
International Institute of Information Technology, Hyderabad GPA: 9.10/10

Work Experience

2024–
Present **Research Intern : CHAI, UC Berkeley, USA**
Working with [Dr. Stuart Russell](#) and [Benjamin Plaut](#) on LLM uncertainty quantification methods for complex multi-turn settings

2024–2025 **Research Assistant: Precog (IIIT Hyderabad), Hyderabad, India**
Multimodal Adversarial Attacks: (Collaboration with Microsoft Research) Finding bounds for adversarial attack transferability between modalities.
LLM Reasoning with World Knowledge Conflicts: Investigating how LLM performance on tasks varies with respect to their parametric memory of the world.

2024 **Machine Learning Product Engineering Intern : Sprinklr Gurgaon, India**
Developed a custom LLM-based Multi-Agent framework (eg: [Autogen](#)) for easy understanding and analysis of structured data. Compared to off-the-shelf frameworks, the framework delivers high accuracy (+30%) on complex tasks along with low latency (85% less) in a cost-efficient (80% less) manner while speeding up the time taken on sales analysis by 28x compared to current human efforts.

2024 **Teaching Assistant, IIITH : Responsible and Safe AI Systems**

2024 **Facilitator : AI Safety Fundamentals - Bluedot Impact (Remote) United Kingdom**
Responsible for facilitating a cohort in the popular [AI Safety fundamentals course \(Course content\)](#).

2023–2024 **Undergraduate Research Assistant : University of Maryland, Baltimore County (UMBC), USA**
(In person)
Evaluating LLMs for reliability and trustworthiness at the Knowledge Infused AI and Inference Lab, led by [Dr. Manas Gaur](#).

2023 **Teaching Assistant, IIITH : Introduction to Natural Language Processing**

2022 **Machine Learning Research Intern : Observe.ai, Bangalore, India**
Developed a novel real-time low-resource multilingual text classification architecture. The framework achieved an F1 score greater than 75 on in-house data, allowing for deployment and enhancing the product's adaptability from English to Spanish users.

2021–2022 **Research Intern : Tokyo Institute of Technology (Remote) Tokyo, Japan**
Worked with [Dr. Kazutoshi Sasahara](#) in curating and statistically analyzing a large corpus of unstructured sensitive data related to Deepfakes, containing 87,000 Reddit comments. (Published at CHI 2022).

2022–2024 **Undergraduate Researcher advised by [Dr. Ponnurangam Kumaraguru](#) : Precog (IIIT Hyderabad), Hyderabad, India**
Codemix-Acceptability (Collaboration with Microsoft Research): Conducted research with [Dr. Monojit Choudhury](#) and [Dr. Manish Shrivastava](#) on creating an acceptability benchmark and

a method to improve generative models for English-Hindi codemix data.

Information Disguise: Worked with [Dr. Joseph Reagle](#) on designing a framework that alters the content of a document (without changing its meaning) and decreases its rank by retrievers like Google Search. (Published at ECIR 2023)

- 2022 **Software Engineer Intern** : *Smart City Living Labs, Hyderabad, India*
Developed and [deployed an immersive and interactive 3D dashboard](#) for real-time monitoring of 300+ IOT Devices (measuring air/water quality, water flow, etc.) across the IIIT Hyderabad campus. Technologies used: Blender, Javascript, ThreeJS, Onem2m.

Papers and Publications

- 2025 **COBIAS: Contextual Reliability in Bias Assessment**
Govil, P., Bonagiri, V., Garg, M., and Kumaraguru, P
ACM Web Science 2025
- 2024 **Evaluating Moral Consistency in Large Language Models**
Bonagiri, V., Vennam, S., Govil, P., Kumaraguru, P., and Garg, M. SaGE
LREC-COLING 20-25 May, 2024, Torino, Italy
- 2024 **K-PERM: Personalized Response Generation Using Dynamic Knowledge Retrieval and Persona-Adaptive Queries**
Raj, K., Roy, K., Bonagiri, V., Thirunarayanan, K
AAAI-MAKE 2024
- 2023 **Representation Learning for Identifying Depression Causes in Social Media**
Govil, P., Bonagiri, V., 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., 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., Whiting, M. E., Sasahara, K
CHI April, 2022
- I Can't Just Assume... Knowledge Conflicts Lead to Bad Reasoning Performance in LLMs**
Bonagiri, V. et al.
(Under review)
- Adversarial Attack Transfers Across Modalities**
Bonagiri, V. et al.
(Under review)
- Acceptability of Code-Mixed Data**
Kodali, P., Goel, A., Bonagiri, V.
(Under review at a journal)

Honours, Awards, and Programmes

- 2024 Dean's List 2 for Spring Semester (Top 10%)

2022	Dean's List 2 for Spring Semester
2021	Dean's List 1 for Monsoon Semester
2021	Merit List for Spring Semester
2021	Winner, Megathon 2021 : Built a Mask Detector to track the status of people's masks (on/off) in real-time in crowds (during Covid).
2020	Merit List for Monsoon Semester
2017	Winner, NASA AMES Space Settlement Contest 2017 : Ranked Top 3 internationally for developing Ingress: A space colony prototype for space colonization.

Other positions of interest

2021-2023	Overall Co-ordinator of Open Source Developers group , IIIT Hyd
2020-2022	Tech Team member, Entrepreneurship Cell , IIIT Hyd
2024-2025	Volunteer, Mental Health Support Group
2020-2025	Freelancer , Have worked on multiple freelance projects such as Blog writing, Book writing and reviews, Poetry Generation, Website creation, Algorithms Teacher, Deep Learning Tutoring etc.