# Ishan Kavathekar

Linkedin ishank31 Google Scholar

Email : ishan.kavathekar@research.iiit.ac.in
Mobile : +91-9370751157

## Education

- **International Institute of Information Technology, Hyderabad** — Hyderabad, IN
  *B.Tech (Hons) and MS by research in Computer Science;* ***CGPA: 8.61*** — *Jul. 2022 – Present*
  - **Honours and Awards**: Deans List (2022, 2023) - Top 10% of the batch
  - **Positions of Responsibility**: Teaching Assistant (TA): Data and Applications, Music Mind and Technology

## Experience

- **Research Intern** — Bangalore, IN
  *Adobe Research* — *May 2025 - Present*
  - Working under Dr. Balaji Vasan Srinivasan in the Multi-Modal Content Group on real-time intelligent suggestions for graphic design, addressing key usability pain points through user-action analysis.
- **Research Intern** — Bangalore, IN
  *Microsoft Research* — *Jan 2025 - May 2025*
  - Collaborating with Tanuja Ganu on evaluating the safety and robustness of LLM agents and multi-agent LLM systems. Work under review at NeurIPs Dataset and Benchmark Track 2025.
- **Undergraduate Researcher** — Hyderabad, IN
  *Precog, IIIT-Hyderabad* — *Apr 2023 - Present*
  - Working under the guidance of Dr. Ponnurangam Kumaraguru. Currently engaged in exploring the risks associated with multi-LLM agent frameworks. Focused on assessing potential vulnerabilities and ethical concerns to ensure the development of safe and reliable AI systems. Addtionally working on small language models (SLMs) for function calling.

## Publications

- **Kavathekar, I.**, Rani, A., Chamoli, A., Kumaraguru, P., Sheth, A., Das, A. (2024). Counter Turing Test ($CT^2$): Investigating AI-Generated Text Detection for Hindi–Ranking LLMs based on Hindi AI Detectability Index ($ADI_{hi}$). **EMNLP 2024 Findings**
- **Kavathekar, I.**, Donakanti, R., Kumaraguru, P., Vaidhyanathan, K. (2025). Small Models, Big Tasks: An Exploratory Empirical Study on Small Language Models for Function Calling. **EASE 2025 AI Models and Data Evaluation Track**.
- **Kavathekar, I.**, Jain, H., Rathod, A., Kumaraguru, P., Ganu, T. (2025). TAMAS: A Dataset for Investigating Security Risks in Multi-Agent LLM Systems. **ICML MAS Workshop 2025.**
- Tripathi, Y., Donakanti, R., Girhepuje, S., **Kavathekar, I.**, Vedula, B. H., Krishnan, G. S., ... Kumaraguru, P. (2024). InSaAF: Incorporating Safety through Accuracy and Fairness | Are LLMs ready for the Indian Legal Domain? **JURIX 2024**

## Projects

- **Adversarial evaluation of LIME for Hindi text**: Adapted the XAIFooler attack method for Hindi by developing a sequential perturbation algorithm that generates adversarial explanations while preserving semantic integrity and prediction stability. Fine-tuned IndicBERT and XLM-RoBERTa models on Hindi datasets, showcasing the applicability of adversarial techniques to low-resource languages.
- **Neural POS Tagger**: Developed a neural POS tagger employing feedforward neural networks and LSTM models, achieving an impressive 98% accuracy for both architectures.
- **ELMo-Based Text Classification System**: Implemented an ELMo (Embeddings from Language Models) architecture from scratch using PyTorch, including a stacked Bi-LSTM network for contextual word embeddings. Pre-trained the model on a bidirectional language modeling task and fine-tuned it for a classification task.

## Programming Skills

- **Languages**: Python, C/C++, mySQL, HTML, CSS, Javascript
- **Frameworks**: Hugging Face, Scikit-learn, PyTorch — **Technologies**: Postman, Git, OpenAI API