Ishan Kavathekar

Linkedin ishank31 Google Scholar

Email : ishan.kavathekar@research.iiit.ac.in
ishankavathekar31@gmail.com
Mobile : +91-9370751157

Education

- **International Institute of Information Technology, Hyderabad** Hyderabad, IN
  B.Tech (Hons.) and MS by Research in Computer Science and Engineering ***CGPA: 8.63*** *Jul. 2022 — Expected 2026*
  - **Advisor**: Dr. Ponnurangam Kumaraguru

Experience

- **Research Intern** Bangalore, IN
  Adobe Research *May 2025 — Aug 2025*
  - Supervisor: Dr. Balaji Vasan Srinivasan
  - Worked on providing real-time intelligent design suggestions for graphic design based on user actions.
  - Designed and implemented a Vision-Language Model (VLM) fine-tuning pipeline, training models on over 600K designs to generate actionable design suggestions based on user actions.
  - A patent has been filed based on the work (currently under review).

- **Research Intern** Bangalore, IN
  Microsoft Research *Jan 2025 — May 2025*
  - Supervisor: Tanuja Ganu
  - Investigated robustness of multi-agent LLM frameworks to adversarial prompting, identifying the vulnerabilities in common multi-agent design patterns. Explored potential defense strategies.
  - Developed SafeAgents, a unified framework for fine-grained safety assessment of multi-agent systems.
  - Work under review at ICLR 2026.

- **Student Research Collaborator** Remote
  Artificial Intelligence Institute, University of South Carolina *Jul 2023 — Aug 2024*
  - Collaborated with Dr. Amitava Das on evaluating the generalizability of five state-of-the-art AI-generated text detection methods for Hindi language.
  - Evaluated various LLMs across five state-of-the-art detection methods and developed AG_{hi} dataset consisting of 36K AI-generated and human-written news articles. Introduced Hindi AI-Detectability Index (ADI_{hi}) to evaluate a model's detectability.
  - Work accepted at EMNLP 2024 Findings.

- **Undergraduate Researcher** Hyderabad, IN
  Precog, IIIT-Hyderabad *Apr 2023 — Present*
  - Working under the guidance of Dr. Ponnurangam Kumaraguru. Currently engaged in exploring the safety and reliability of LLM-agents and multi-agent LLM systems under adversarial settings.
  - Focused on evaluating capabilities, potential vulnerabilities, and safety concerns in LLMs to ensure the development of safe and reliable AI systems.

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** [pdf]

- **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** [pdf].

- **Kavathekar, I.**, Jain, H., Rathod, A., Kumaraguru, P., Ganu, T. (2025). TAMAS: Benchmarking Adversarial Risks in Multi-Agent LLM Systems. **ICML MAS Workshop 2025. Under Review.** [ICML-W] [pdf]
- Tripathi, Y., Donakanti, R., Girhepuje, S., **Kavathekar, I.**, Vedula, B. H., Krishnan, G., Goyal, S., Goel, A., Ravindran, B., Kumaraguru, P. (2024). InSaAF: Incorporating Safety through Accuracy and Fairness — Are LLMs ready for the Indian Legal Domain? **JURIX 2024** [pdf]
- Arora, N., Joel, S., **Kavathekar, I.**, LNU, P., Gandhi, R., Pandya, Y., Ganu, T., Kanade, A., Nambi, A. (2025). Exposing Weak Links in Multi-Agent Systems under Adversarial Prompting. **Under Review.** [pdf]

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 accuracy of 98% for both architectures. [code]
- **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. [code]
- **C-Shell**: Developed a bash-like command interpreter using C language and system calls. Implemented built-in commands (cd, echo, pwd, ls) and advanced features such as foreground/background process execution, along with input and output redirection. [code]

Technical Skills

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

Teaching Roles

- **Teaching Assistant - Data and Applications (Fall 2023, Fall 2025)**:
  Designed assignments, conducted weekly tutorials, graded exams and assignments, and mentored student projects.
- **Teaching Assistant - Music, Mind and Technology (Spring 2024)**:
  Designed assignments, conducted weekly tutorials, graded exams and assignments, and mentored student projects.
- **Teaching Assistant - Learning and Memory (Fall 2024)**:
  Graded exams and assignments, and mentored student projects.

Honours and Awards

- **AI Evaluation Programme - Jan 2026**:
  Selected as one of 40 students worldwide for a fully funded program focused on AI evaluation for capabilities and safety
- **Deans List - (Spring 2023, Spring 2024, Fall 2024)**:
  Recognized for academic excellence as one of the top performers in the academic year.
- **Research Award - 2025**:
  Recognized for publishing at an international conference as an undergraduate student.
- **First Place - Megathon - 2022**:
  Secured first place in Qualcomm's Accent Detection track at Megathon 2022 hackathon.

Academic Service

- **Reviewer** - ICML MAS Workshop 2025
- **Reviewer** - ACM IKDD CODS 2025

Relevant Coursework

Statistical Methods in AI, Advanced NLP, Introduction to NLP, Optimization Methods, Advanced Computer Networks, Design and Analysis of Software System, Behavioral Research Statistical Methods, Algorithm Analysis and Design, Performance Modeling of Computer Systems (Queuing Theory), Introduction to Information Security, Operating Systems & Networks, Probability & Statistics.