Shanu Kumar

PhD Student at Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
✉ Shanu.kumar@mbzuai.ac.ae | sshanukr@gmail.com | Google Scholar | Website

Education

2019 **B. Tech in Electrical Engineering**, Indian Institute of Technology, Kanpur — GPA: 8.63/10
Silver Medalist, Minor: Machine Learning
2015 **Central Board of Secondary Education**, Adarsh Vikash Vidaylaya, Patna — GPA: 92.4 %

Research Interests

Personalization, Alignment, Spatial Reasoning, and Interpretability in Vision–Language Models

Achievements

Microsoft Awards
2025 Won two awards at the **IDC Innovator Arena** for auto-prompt tuning: **Most Innovative Solution** & **People's Choice**.
2024 **STCI Excellence Award** for envisioning prompt auto-tuning.
2024 **Spot Award** for developing a compact language model for content moderation.
2023 **STCI Excellence Award** for advancing prompt engineering techniques for LLMs.
2023 **1st & 3rd Place Finishes** in company-wide "Aspire Hack" and "Executive Challenge" hackathons.
2022 **Spot Award** for creating a multilingual content moderation system.

Academic Honors & Grants
2019 **Proficiency Prize** from IIT Kanpur for exceptional undergraduate research.
2019 **Research Grants** from Microsoft Research (x2) and the Indian National Academy of Engineering (INAE).
2017 **Academic Excellence Award** for ranking in the top 5% of students at IIT Kanpur.
2016 **MCM Scholarship** for sustained academic excellence at IIT Kanpur.
2015 Secured **All India Rank 2499** in JEE Advanced (Top 2% of over 125,000 candidates).

Publications & Patents

Patents

2025 Systematic Tuning of Prompt Systems
Shanu Kumar, *Akhila Yesantarao Venkata, Shubhanshu Khandelwal, Parag Agrawal, Manish Gupta*
Patent Application (Under Review)

2023 Neural-Aided Clique Graph Mining for Low Authority Host and URL Detection
Shanu Kumar, *Sai Krishna Mendu, Avinash Kumar*
Patent (Granted)

Peer-Reviewed Publications

2025 LITMUS++: An Agentic System for Predictive Analysis of Low-Resource Languages Across Tasks and Models
Avni Mital Kumar, **Shanu Kumar**, *Sandipan Dandapat, Monojit Choudhury*
AACL Demo Paper 2025

2025 SCULPT: Systematic Tuning of Long Prompts
Shanu Kumar, *Akhila Yesantarao Venkata, Shubhanshu Khandelwal, Bishal Santra, et al.*
ACL Mains 2025 [Paper]

2025 Towards Safer Pretraining: Analyzing and Filtering Harmful Content in Webscale datasets
Sai Krishna Mendu, Harish Yenala, Aditi Gulati, **Shanu Kumar**, *Parag Agrawal*
IJCAI Mains 2025 [Paper]

2025 Navigating the Cultural Kaleidoscope: A Hitchhiker's Guide to Sensitivity in LLMs
Somnath Banerjee, Sayan Layek, Hari Shrawgi, Rajarshi Mandal, Avik Halder, **Shanu Kumar**, *et al.*
NAACL Mains 2025 [Paper]

2025 SafeInfer: Context Adaptive Decoding Time Safety Alignment for LLMs
Somnath Banerjee, Soham Tripathy, Sayan Layek, **Shanu Kumar**, *Animesh Mukherjee, Rima Hazra*
AAAI 2025 [Paper]

2025 Enhancing Zero-shot CoT Prompting via Uncertainty-Guided Strategy Selection
Shanu Kumar, *Saish Mendke, Karody Lubna Abdul Rahman, Santosh Kurasa, et al.*
COLING Oral 2025 [Paper]

2025 Socio-Culturally Aware Evaluation Framework for LLM-Based Content Moderation
Shanu Kumar, *Gauri Kholkar, Saish Mendke, Anubhav Sadana, Parag Agrawal, Sandipan Dandapat*
SUMEval Workshop @ COLING 2025 [Paper]

2023 DiTTO: A Feature Representation Imitation Approach for improving Cross-Lingual Transfer
Shanu Kumar, *Soujanya Abbaraju, Sunayana Sitaram, Sandipan Dandapat, Monojit Choudhury*
EACL Mains 2023 [Paper]

2022 "Diversity and Uncertainty in Moderation" are the Key to Data Selection...
Shanu Kumar, *Sandipan Dandapat, Monojit Choudhury*
NAACL Findings 2022 [Paper]

2022 Multi Task Learning For Zero Shot Performance Prediction of Multilingual Models
Kabir Ahuja,* **Shanu Kumar***, *Sandipan Dandapat, Monojit Choudhury*
ACL Oral 2022 [Paper]

2019 Attending to Discriminative Certainty for Domain Adaptation
Vinod Kumar Kurmi,* **Shanu Kumar***, *Vinay P. Namboodiri*
CVPR 2019 [Paper]

2019 Adversarial Adaptation of Scene Graph Models for Understanding Civic Issues
Shanu Kumar, *Shubham Atreja, Anjali Singh, Mohit Jain*
WWW 2019 [Paper]

Preprints & Under Review

2025 UMOPRO: Uncertainty-Aware Multi-Objective Prompt Optimization
Shanu Kumar, *Shubhanshu Khandelwal, Akhila Yesantarao Venkata, et al.*
Under Review

2025 Attributional Safety Failures in LLMs under Code-Mixed Perturbations
Somnath Banerjee, Pratyush Chatterjee, **Shanu Kumar**, *Sayan Layek, Parag Agrawal, et al.*
Under Review [Paper]

2023 READ: Reinforcement-based Adversarial Learning for Text Classification...
Rohit Sharma,* **Shanu Kumar***, *Avinash Kumar*
Preprint [Paper]

2020 Mitigating Uncertainty of Classifier for Unsupervised Domain Adaptation
Shanu Kumar, *Vinod Kumar Kurmi, Praphul Singh, Vinay P. Namboodiri*
Preprint [Paper]

Work Experience

Microsoft, Data & Applied Scientist (2019 – 2025)

2025
2024 **Automatic Prompt Engineering & Optimization**

- **Architected SCULPT**, an automatic prompt optimization framework using targeted edits and aggregated feedback. Deployed across dozens of product teams, driving major gains (**up to 10%** in products like Copilot).
- Led the integration of SCULPT into **centralized model development and data labeling platforms** to standardize prompt tuning company-wide. Published the work at **ACL 2025** and filed a **U.S. patent**.
- Designed a novel Pareto-front algorithm to jointly optimize prompts for both **performance and efficiency**, enabling tailored latency/quality trade-offs.
- Led evaluation and enhancement of critical safety prompts, improving **jailbreak detection** through manual and automated tuning.

2024
2023 **Content Moderation Systems & LLM Migration**

- **Designed a new architecture for the core content moderation model**, scaling it to new temporal domains (e.g., Elections) and achieving **double-digit improvements in recall**.
- **Led prompt migration from legacy systems to next-generation LLMs**. Improved the model's F1-score and precision on key tasks while **significantly reducing prompt length**.
- **Fine-tuned a Small Language Model (SLM)** using QLoRA that **outperformed GPT-4** on key labeling tasks, enabling data generation at a scale of **millions of samples**.
- Engineered a state-of-the-art safety prompt with hundreds of Chain-of-Thought examples, improving safety classification recall by **over 25%**.

2023
2022 | Foundational Safety Models & Data Pipelines

- **Shipped the foundational unified risk model**, consolidating multiple legacy classifiers. This single model blocked **millions of additional harmful suggestions** and reduced over-triggering by **over 10%**.
- **Developed and shipped a 3-layer distilled model** for online, low-latency use cases, improving recall by **over 20%** while reducing latency by **over a third**.
- Created robust evaluation "goldsets" using novel **uncertainty** and **agreement/disagreement sampling** techniques to systematically uncover model vulnerabilities.
- Developed a universal threat model to detect **over 90 distinct harm types** in a multilingual setting.

2022
2020 | Multilingual & Multi-Modal Classification

- **Shipped a universal model for Adult Intent classification in over two dozen languages** with consistent cross-lingual performance.
- Developed a multi-modal document classifier leveraging both HTML text and images.
- Created a novel algorithm using Data Cross-Entropy to improve the precision of false-negative query retrieval by **nearly 30%**.
- Implemented a meta-learning framework for label correction to handle noisy data in few-shot settings.

2019 | Unsupervised Question Answering Systems

- Architected an encoder-decoder model to automatically extract relevant query-passage pairs from web documents for technical troubleshooting scenarios.
- Pioneered research into applying open-domain question answering techniques in a zero-shot, unsupervised setting to solve complex support queries.

Internship

2018 | IBM Research India (Research Intern)

- Proposed a novel application of Scene Graph models to generate "Civic Issue Graphs" from images, enabling structured understanding of real-world infrastructure problems.
- Created two novel, multi-modal datasets for civic issue understanding, complete with bounding box annotations and rich text descriptions. Published this work at **WWW 2019**.

Selected Academic & Research Projects

2019 | Semi-Supervised Learning with Deep Generative Models
Course Project in Probabilistic Modeling | *Report*

- Implemented and analyzed two seminal deep generative models, focusing on variational inference methods for semi-supervised classification tasks.

2019 | Unsupervised Domain Adaptation for Semantic Segmentation

Course Project in Visual Recognition | *Report*
- Engineered a progressive domain adaptation pipeline (GTA V → Cityscapes → Custom) to significantly improve segmentation performance on real-world surveillance video.

2019 | **Aligning Classifier Certainty for Domain Adaptation**
Supervisor: Prof. Vinay P. Namboodiri
- Developed a novel method to generate "certainty activation maps" and aligned them across source/target domains to boost classifier confidence and performance.

2019 | **Fine-Grained Classification via Coarse Class Activation**
Course Project in Visual Recognition | *Report*
- Built an end-to-end hierarchical model that improved fine-grained classification by using coarse category probabilities to soft-mask and guide the network's attention.

2018 | **Attending to Discriminative Certainty for Domain Adaptation**
Supervisor: Prof. Vinay P. Namboodiri
- Proposed a novel attention mechanism that identifies adaptable regions in an image based on the certainty estimates of a discriminator.
- **Achieved state-of-the-art results** on three benchmark datasets: Office-Home, Office-31, and ImageCLEF-2014.

2018 | **Mining and Prediction of Civic Issues**
Course Project in Data Mining | *Report*
- Designed a system to automatically categorize and assign civic issue complaints (e.g., potholes, sanitation) using titles, descriptions, and images.

2018 | **Bayesian Neural Networks for Domain Adaptation**
Supervisor: Prof. Vinay P. Namboodiri
- Formulated a Bayesian framework for domain adaptation by transforming the classifier and discriminator into Bayesian NNs using Monte Carlo Dropout for uncertainty estimation.

2018 | **Hierarchical Word Sense Disambiguation**
Supervisor: Prof. Harish Karnick | *Report*
- Developed an end-to-end hierarchical model using CNNs and WordNet senses to sequentially predict the correct sense for each word in a sentence.

2018 | **Unsupervised Machine Translation with GCNs**
Course Project in Natural Language Processing | *Report*
- Proposed a Graph Convolutional Network (GCN) based autoencoder to impose grammatical structure onto the latent space for unsupervised machine translation.

2018 | **Visual Motor Control of a Robotic Arm**
Course Project in Neural Networks | *GitHub*
- Implemented Single Network Adaptive Critic (SNAC) and Self-Organizing Maps (SOM) in TensorFlow for the visual motor control of a multi-joint robotic arm.

2017 | **Bidirectional Attention Flow for Machine Comprehension**
Course Project in Machine Learning | *Report*
- Implemented the BiDAF model, a foundational architecture for question answering, and explored enhancements using grammatical features like part-of-speech embeddings.

2017 | **Relation Classification using Tree LSTMs**
Supervisor: Prof. Harish Karnick | *GitHub*
- Developed a model using Bidirectional Tree LSTMs on dependency paths to classify the semantic relation between two entities in a sentence.

Hackathons

2024 | **InstaDiagram**, Microsoft Global Hackathon

- Created an AI tool that converts descriptions or sketches into professional, editable diagrams instantly, boosting productivity and visualization in system design and data science.

2023 **SHARE TO UPSKILL**, Microsoft Global Hackathon

- Created a dynamic platform for peers to share and develop personal skills, encompassing a diverse range of cultural, technical, and mental well-being competencies.

2023 **PROJECT MATE**, Microsoft LLM Hackathon

- Developed a platform that offers peers a dynamic platform to share and cultivate personal skills, spanning a rich tapestry of cultural, technical, and mental well-being competencies.

2022 **BIAS EVALUATION TOOL**, Microsoft Global Hackathon

- Developed a tool to identify the biases present in AI models and deep dive into what exactly is causing the unwanted biases in the model. .

2017 **QALEARN**, Microsoft Code.Fun.Do.

github

- Developed a Web Application for open-domain question answering on ebooks using BiDAF model.

2016 **AUTOMATED LIBRARY**, Microsoft Code.Fun.Do.

github

- Developed a Web Application in Django to catalogue bibliographies and library members.