
SUMMARY

AI researcher at Sony Research, working in causal inference for user engagement modeling. Previously, worked at IIT Delhi, on LLM multi-agent systems, in collaboration with Microsoft Research and DRDO. Published at EMNLP, EACL (Oral), AACL, TMLR, IJCNN with hands-on experience shipping RAG pipelines, and multi-modal systems to production.

SKILLS AND INTERESTS

Interests	Applications of LLM, Trustworthy AI, Generative AI, Multimodality, Agentic Systems
Programming	Python, C, SQL
Frameworks	PyTorch, TensorFlow, Unsloth, FastAPI, Docker, LangChain, LlamaIndex, Agentic Orchestration (LangGraph, CrewAI, Autogen), LLM APIs (Azure OpenAI, Google Vertex AI, Huggingface open-source LLMs), Vector DBs, Cloud (Azure and AWS basic), git
Concepts	Multi-agent Orchestration, Tool Calling, Memory Management, Retrieval-Augmented Generation (RAG), LLM for MRG, Hallucination Mitigation, Prompt Engineering (COT, TOT, ReAct), LLM Inference Optimization (vllm), Fine-tuning (SFT, PEFT, LORA), Quantization

EXPERIENCE

- Sony Research India** Jun 2026 - Present
Research Fellow Bengaluru, India (Remote)
 - Working in the User Engagement Research group. Primarily, focusing on modeling user behavior of the SonyLiv platform using causal inference, and creating an agentic simulation platform.
- Indian Institute of Technology (IIT) Delhi, Laboratory for Computational Social Systems** Jul 2025 - May 2026
Research Associate Delhi, India (Onsite)
 - Worked on an Agentic AI project in collaboration with **Microsoft Research India (MSRI)**, co-supervised by Prof. Tanmoy Chakraborty and principal engineer at MSRI Dr. Akshay Nambi. I proposed a failure-centric evaluation framework for LLM agentic systems, introducing a unified failure taxonomy to identify execution-level breakdowns across an agent's reasoning, planning, and multi-step decision pipelines. Analyzed multi-agent orchestration frameworks (Autogen, CrewAI, LangChain) to diagnose planning stability, tool calling, and reasoning-action alignment, optimizing reliability for agentic workflows beyond simple task success rates.
 - Collaborated with **DRDO** to develop INFORM, an interpretability framework for multi-expert LLM orchestration. Conducted analyses on GSM8K, HumanEval, and MMLU using LLaMA-3.1 8B, Qwen3 8B, and DeepSeek-R1 8B expert ensembles. Combined gradient sensitivity analysis with interaction topology to distinguish intrinsic expert importance from routing behavior, demonstrating that frequently selected experts are not necessarily the most functionally critical. *Accepted at TMLR 2026.*
 - Proposed FIRE, a multi-agent framework for category-aware, fact-grounded counterspeech generation, alongside FactualCS, a 4,784 instance annotated dataset with reasoning traces and evidence mappings. Demonstrated SOTA performance over 28 baselines, achieving ~ 12% higher factual accuracy, ~ 11% better category-specific accuracy, and ~ 11% lower toxicity while leveraging compact (<2B) language models. *Accepted at EMNLP 2026 Mains.*
- Indian Statistical Institute (ISI), Electronics and Communication Sciences Unit (ECSU)** June 2024 - June 2025
Research Collaborator supervised by Prof. Swagatam Das Kolkata, India (Onsite)
 - Contributed to ARREST (Adversarial Resilient Regulation Enhancing Safety and Truth), a unified framework to mitigate factual and safety failures in LLMs by addressing representational misalignments in the latent activation space. Engineered an external intervention network to self-correct drifted features, regulating falsehoods and unsafe outputs without fine-tuning model parameters, demonstrating superior versatility over standard RLHF models in handling soft refusals. *Accepted at EACL 2026 (Oral).*
 - Contributed to HalluShift, a novel evaluation pipeline for detecting factual hallucinations and ensuring model safety by analyzing distributional shifts in the internal state space and token probabilities of LLM-generated responses. *Accepted at IJCNN 2025.*
 - Developed a Medical Report Generation (MRG) system by leveraging Gemini-1.5-Flash to create Simplified Medical Reports (SMRs). Employed advanced prompt engineering strategies (Chained Prompting, In-Context Learning) to structure structured outputs and batch API queries. Fine-tuned open-source LLaMa models on generated SMRs, reducing inference cost while maintaining clinical accuracy. *Accepted as Student Abstract at AACL 2025.*
 - Built an image captioning pipeline using BLIP for embedding generation and CerberusDet (YOLOv8) for object detection. Formatted data in Alpaca-style and fine-tuned multiple LLMs (LLaMa 3.1, LLaMa 2, Mistral 7B, Phi-2) to generate structured outputs. Evaluated generative models using BLEU, ROUGE, and BERT-based metrics.
- Defence Research and Development Organisation (DRDO), DEBEL Lab** March 2024 - April 2024
Summer Internship Bengaluru, India (Onsite)
 - Developed comprehensive dataset using IMU sensor data from Xsens, capturing walking patterns of 20 subjects. Applied computer vision and machine learning techniques for motion analysis.

- Conducted research on human gait pattern analysis using ensemble learning methods, implementing real-time prediction systems for lower limb prosthetics.

• Geogo Techsolutions

October 2023 - May 2025

Machine Learning Engineer Intern

Kolkata, India (Hybrid)

- Co-led the development of Kriyam DocWise, an AI-powered document assistant that enables context-aware, natural-language access to policy, claim and ID documents. Architected a scalable hybrid Retrieval-Augmented Generation (RAG) pipeline and built REST APIs using FastAPI for backend deployment, integrating OCR, vector embeddings, similarity search, and summarization to eliminate manual review and accelerate policy analysis for claims investigation teams.
- Built an end-to-end face comparison and similarity search solution, achieving 98% accuracy in benchmarks against Amazon Rekognition. Implemented state-of-the-art computer vision models for employee authorization and face verification, successfully matching current faces with historical photos. Deployed with kriyam.ai.
- Developed a liveness detection system to flag pre-recorded videos and combat insurance fraud using hand/eye movement analysis and 3D facial mesh inspection. Beta-tested across 5,000 users in video e-KYC verification.
- Engineered an Automatic Speech Recognition (ASR) model using Transformer architecture, trained on a 3-hour Hinglish (Hindi+English) dataset. Achieved 34% WER, now deployed for insurance call transcription.

RESEARCH PUBLICATIONS

* indicated authors contributed Equally to the work

1. **Sujoy Nath***, Aswini Kumar*, Tanmoy Chakraborty, "Counter with Evidence! A Multi-Agent Memory Efficient Reasoning Framework for Hate Category Informed Counterspeech Generation", Accepted, **EMNLP 2026 Mains**, Ranking: **A***, <https://arxiv.org/abs/2608.23152>
2. Sudipto Ghosh*, **Sujoy Nath***, Sunny Manchanda, Tanmoy Chakraborty, "Disentangling Intrinsic Importance from Emergent Structure in Multi-Expert Orchestration", **Transactions on Machine Learning Research (TMLR) 2026**, <https://openreview.net/forum?id=4W7sgat04A>
3. Sharanya Dasgupta, Arkaprabha Basu, **Sujoy Nath**, Swagatam Das, "ARREST: Adversarial Resilient Regulation Enhancing Safety and Truth in Large Language Models", **EACL 2026 (Oral)**, Ranking: **A**, <https://aclanthology.org/2026.eacl-long.212/>
4. **Sujoy Nath**, Arkaprabha Basu, Kushal Bose, Swagatam Das, "From Complexity to Clarity: Transforming Chest X-ray Reports with Chained Prompting", **AAAI 2025 Student Abstract**, Ranking: **A***, <https://ojs.aaai.org/index.php/AAAI/article/view/35281>
5. Sharanya Dasgupta, **Sujoy Nath**, Arkaprabha Basu, Pourya Shamsolmoali, Swagatam Das, "HalluShift: Measuring Distribution Shifts towards Hallucination Detection in LLMs", **IJCNN 2025**, Ranking: **B**, <https://ieeexplore.ieee.org/document/11228484>
6. **Sujoy Nath**, Arkaprabha Basu, Sharanya Dasgupta, Swagatam Das, "HalluShift++: Bridging Language and Vision through Internal Representation Shifts for Hierarchical Hallucinations in MLLMs", **ICVGIP 2025 (Oral)**, <https://dl.acm.org/doi/10.1145/3774521.3774542>

EDUCATION

• Maulana Abul Kalam Azad University of Technology

August 2021 - June 2025

Bachelor of Technology in Computer Science and Business System, CGPA: 8.66/10

Kolkata, India

ACHIEVEMENTS

• Parayas 2k24: National Level Project Competition

May, 2024

2nd position, Software segment (Team Lead)

Kolkata, India

• Kavach 2023: National Cybersecurity Hackathon

Aug, 2023

Top 5 finalist teams (Team Lead), Ministry Of Education, Govt of India

Greater Noida, India

RELEVANT COURSEWORK

Machine Learning (Grade: 16/20), Artificial Intelligence (Grade: 27/30), Pattern Recognition (Grade: 24/30), Programming for Problem Solving (Grade: 30/30)