SUJOY NATH

Email : sujoynathofficial@gmail.com
sujoynath.in, Google Scholar, Linkedin, Github

SUMMARY

AI researcher at IIT Delhi, working on LLM multi-agent systems, and LLM safety in collaboration with Microsoft Research India and DRDO. Published at EACL (Oral), AAAI, IJCNN with hands-on experience shipping RAG pipelines, and computer vision 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, Fine-tuning (SFT, PEFT), Quantization (LoRA)

RESEARCH EXPERIENCE

- **Indian Institute of Technology (IIT) Delhi, Laboratory for Computational Social Systems** *Jul 2025 - Present*
  Research Associate Delhi, India (Onsite)
  - Working 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 agentic systems powered by large language models, 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. (Paper under review)
  - Collaborated with **DRDO** to introduce INFORM, an interpretability analysis treating multi-expert system orchestration as explicit computation. Evaluated LLM orchestrators on GSM8K, HumanEval, and MMLU using LLaMA-3.1 8B, Qwen3 8B, and DeepSeek-R1 8B. Applied gradient-based causal attribution to disentangle expert interaction structure from execution order, revealing that an expert's intrinsic structural importance diverges significantly from mere routing frequency.

- **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 AAAI 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.

Industry Experience

- **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. 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/
2. **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
3. 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
4. **Sujoy Nath**, Arkaprabha Basu, Sharanya Dasgupta, Swagatam Das, *"HalluShift++: Bridging Language and Vision through Internal Representation Shifts for Hierarchical Hallucinations in MLLMs", Accepted*, **ICVGIP 2025 (Oral)**, https://arxiv.org/abs/2512.07687

Manuscripts Under Review:

5. Sudipto Ghosh*, **Sujoy Nath***, Sunny Manchanda, Tanmoy Chakraborty, *"Disentangling Causal Importance from Emergent Structure in Multi-Expert Orchestration"*, Under Review (preprint) at **Transactions on Machine Learning Research (TMLR)**, https://arxiv.org/abs/2602.04291
6. **Sujoy Nath**, Aswini Kumar, Archana Yadav, Akshay Nambi, Tanmoy Chakraborty, *"When Do Agents Go Wrong: A Failure-Centric Evaluation of Agentic Systems"*, Under Review
7. **Sujoy Nath***, Aswini Kumar*, Tanmoy Chakraborty, *"Counter with Evidence! A Multi-Agent Memory Efficient Reasoning Framework for Hate Category Informed Counterspeech Generation"*, Under Review

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)