Riju Marwah

New Delhi, India | marwah.riju@gmail.com | linkedin.com/in/riju-marwah | rijumarwah.github.io

Research Interests: Trustworthy & Reliable LLMs, Token-level Dynamics, Interpretability, Alignment

Profile

Undergraduate Computer Science researcher with interests in trustworthy and reliable large language models, mechanistic interpretability, and token-level dynamics. I am currently a **Research Intern at the IRT Group, University of South Carolina**, supervised by computer scientist **Dr. Amit Sheth**. I have authored and presented work at Core A* Conferences like AAAI and ICML.

Education

Bachelor of Technology | ***Computer Science and Engineering*** 11/2022 – 06/2026 (Expected)
Guru Gobind Singh Indraprastha University, India **GPA: 8.2/10.0**
Relevant Coursework: Artificial Intelligence, Data Structures & Algorithms, Operating Systems, Computer Networks, Compiler Design

Work Experience

Research Intern 04/2025 – Present
Artificial Intelligence Institute, University of South Carolina Columbia, SC, USA

- **Principal Investigator:** Dr. Amit Sheth (NCR Chair & Director, AIISC), **Advisor:** Vishal Pallagani
- Designed a framework to detect and mitigate cognitive fatigue in LLMs using token-level signals and real-time interventions.
- First-authored accepted papers at AAAI and ICML 2026.
- Conducting research on long-context reliability, entropy collapse, and attention decay in LLMs.

Research Collaborator 11/2025 – 01/2026
University of Illinois Urbana–Champaign Champaign, IL, USA

- **Advisor:** Soorya Ram Shemgekar
- Studying politeness framing and reward leakage in LLMs across structured tasks and instruction-following settings.
- Performed mechanistic interpretability analysis including early-token probing and activation patching.

Generative AI Intern 01/2025 – 03/2025
EY (Ernst & Young) New Delhi, India

- Built a low-code platform for agentic AI workflows using modular DAGs, vector DBs, and LLM toolchains.
- Implemented Celery-Redis task execution with production-grade scalability.
- Integrated semantic agent routing, memory components, and external API/tool support.

Software Developer Intern 07/2024 – 09/2024
National Thermal Power Corporation New Delhi, India

- Developed ASP.NET Core applications using MVC and Entity Framework for enterprise automation.
- Optimized MySQL and MongoDB CRUD operations via LINQ, ensuring ACID compliance.
- Implemented secure authentication using JWT, Identity Framework, Google reCAPTCHA, and SMTP.

Research Publications

Cognitive Fatigue in Autoregressive Transformers: Formalization and Measurement
ICML 2026 (Main Track, First Author)

- Formalized cognitive fatigue in autoregressive LLMs and introduced the Fatigue Index (FI), a lightweight model-agnostic diagnostic aggregating attention decay, representational drift, and entropy miscalibration.
- Evaluated FI across nine models (1B–13B), predicting task degradation (AUROC = 0.95) and repetition ($p = 0.94$) with stress analyses revealing sensitivity to context length, evidence position, and numerical precision.

Chatsparent: An Interactive System for Detecting and Mitigating Cognitive Fatigue in LLMs
AAAI 2026 (Demonstration, First Author)

- Built a live diagnostic interface that streams the Fatigue Index alongside model outputs, making generation reliability visible and measurable in real time.

- Implemented retrain-free interventions triggered on threshold crossings, restoring generation stability without modifying model weights.

MicroDetect-Net (MDN): Leveraging Deep Learning to Detect Microplastics in Clam Blood
Springer Nature, ICICC-2025

- Developed an AI tool for detecting microplastics in blood samples, addressing a key environmental health challenge.
- Applied RGB thresholding and binary masking to isolate microplastics in fluorescence microscopy images.

PROJECTS

Early Stage Lung Cancer Detection Using Deep Convolutional Neural Networks
Built a CNN model using the LIDC-IDRI dataset to detect pulmonary nodules in CT scans, enabling earlier lung cancer diagnosis. Integrated explainability tools (Grad-CAM, SHAP) to assist radiologists in making informed decisions.

Bhoomi, AI Powered Agricultural Productivity
Developed an AI-powered full-stack platform to enhance agricultural productivity, integrating drone technology and mobile solutions for better yield and sustainability, developing AI models for crop disease detection and farmer assistance.

VOLUNTEERING

Student Scholar Volunteer, Association for the Advancement of Artificial Intelligence (AAAI) 2026
Volunteering at the AAAI 2026 Conference in Singapore, while presenting an accepted paper.

Delegate, Harvard Project for Asian & International Relations 2025
Selected for the prestigious HPAIR Conference.

Volunteer, Rotaract Club of New Delhi 2025
Led community health initiatives in Delhi's urban slums.

HONORS

Awarded the AAAI Student Volunteer Scholarship
competitively selected for USD 1,100 travel grant to present at AAAI 2026 in Singapore.

Featured in the News
Times of India, Dwarka Parichay, Brainfeed (Quarantech), Hindustan Times (Ordin@trix).

Selected for the McKinsey Forward Program
A global initiative on problem-solving, business, and leadership.

Presented at International Conferences on AI
At ICICC-2025, and AAAI-2026

ADDITIONAL QUALIFICATIONS

Massachusetts Institute of Technology, OpenCourseWare, Graduate Coursework/Self-Study
6.825 – Techniques in Artificial Intelligence/Deep Learning for NLP
6.864 – Advanced Natural Language Processing
6.867 – Machine Learning (Advanced)
6.006 – Introduction to Algorithms

SKILLS

Programming: Python, C++
AI/ML: PyTorch, TensorFlow, Hugging Face, Scikit-learn, LangChain, LIME, SHAP
Backend & APIs: Flask, FastAPI, REST, JWT
Databases: SQL (MySQL, PostgreSQL), NoSQL (MongoDB, Redis)
Misc: LaTeX, Markdown

LANGUAGES

English: Native — TOEFL iBT: 107/120
Hindi: Native