RISHABH SABHARWAL

+44 7486045605 ◇ Edinburgh, Scotland

Contact ◇ LinkedIn ◇ Github

OBJECTIVE

I am an **MSc. AI** student at the **University of Edinburgh**. Previously, I worked as a Pre-Doctoral Researcher at the Indian Institute of Science (IISc) under the supervision of Prof. Punit Rathore. My research interests span graph neural networks (GNNs) and large language models (LLMs), with a focus on the Interpretability of LLMs and AI Safety. My prior work includes projects on (a) speech recognition, (b) optimization in computer vision, (c) graph attention networks and spectral GNNs and (d) culturally grounded LLMs.

EDUCATION

MSc. in Artificial Intelligence 2025 - 2026

University of Edinburgh, Scotland

B.Tech. in Electronics and Communication Engineering (AI and ML) 2020 - 2024

Netaji Subhas University of Technology, Delhi, India 8.49/10 or 3.68/4

PUBLICATIONS AND PROJECTS

A.Mazumder, **Rishabh Sabharwal**, M.Tayal, B. Kumar, C. Garg, P. Rathore, 'On Convergence of Adam with Data Dependent Stepsize". (Accepted at **IEEE Transactions on AI**, to be published.)

Rishabh Sabharwal, P.S. Rathore, P. Rathore, "DiGCT: Diffusion Guided Gaussian Context Transformer".

R. Samarth*, **Rishabh Sabharwal***, P. Rathore, "SSGNN: Simple Yet Effective Spectral Graph Neural Network". (Openreview)

Rishabh Sabharwal, S. Syal, P. Pankaj, "Optimizing Stock Market Predictions: A 3-Stage Model with Residual Modelling". In Proceedings of IEEE Signal Processing and Integrated Networks, 2024. (Paper)

FastAPI Safety Gateway: Built a guardrail around a vLLM-served Llama-3.1-8B-Instruct, combining learned and rule-based guardrails like multi-label toxicity/harm classification (toxic-bert), prompt-injection detection (local detector), and PII detection/anonymization (Presidio). Developed an **agentic safety loop** that rewrites risky prompts, repairs unsafe outputs through policy-conditioned re-prompting, and offers user-tunable thresholds, custom deny-lists, and full decision traceability via CLI and Streamlit UI. (Github)

Video Humor Reasoning: Developed the **Indic-SMILE** dataset for generating plausible reasons behind humor in video clips; extracted audio, visual and textual features; benchmarked VLMs and LLMs (InternVL3, VideoLLaMa3 Gemini-2.5 Flash, Claude 4.5 Sonnet, Gemma-3, Sarvam-M) to evaluate cultural alignment in Indian contexts.

Deepfake Image Detection Challenge: An open-source project done with Omdena, where I led the data and model evaluation team. Focused on model's robustness and practicality to images found on social media websites. Achieved 3rd rank in the challenge. (Project Presentation)

EXPERIENCE

Pre-doctoral Researcher June 2024 - July 2025

Indian Institute of Science *Bengaluru, India*

- Worked with **NPCI** (National Payments Corporation of India) to build efficient and scalable large-scale GNNs to detect fraudulent transactions on graphs having more than a **billion nodes**.
- Developed an efficient spectral GNN featuring a novel, parameter-free module called ReGA (Relative Gaussian Amplifier) that performs global context filtering. **Evaluated on 20 real-world graph datasets**; achieved SOTA performance, reducing parameters by an average of 55x and GFLOPs by 100x across all datasets.

- Extended the Gaussian Context Transformer (GCT) to develop DiGCT, utilizing Graph Tikhonov Regularization and graph diffusion for stronger channel attention. Increased average accuracy by **2%** on modern ConvNets with **only 20 additional parameters**.

Research Intern Jan 2024 - June 2024
Indian Institute of Science *Bengaluru, India*

- Proposed a network and data-dependent constant learning rate (lr) to address the issues of decaying lr schedulers and reduce the need for extensive lr search.
- Conducted empirical analysis of our proposed lr which results in faster convergence for both stochastic and deterministic variants of the Adam optimizer. **Paper accepted at IEEE Transactions on AI.**

Research Intern May 2023 - July 2023
Indraprastha Institute of Information Technology *Delhi, India*

- Developed a speech-to-text model to streamline the patient registration process in government hospitals, tackling the problems of **different dialects and frequent language switching** between Hindi and English.
- Fine-tuned and benchmarked Whisper, Meta MMS, VOSK, NVIDIA Nemo models on Indian-accented English and Hindi datasets and some internal medical-speech samples.
- Developed a pipeline that includes a voice activity detector to reduce hallucinations in long audio samples, along with a quantized and fine-tuned Whisper model.
- Our prototype secured **Top-10 rank in the nation-wide Bhashini Grand Innovation Challenge**.

RELEVANT COURSEWORK

- **Accelerated NLP:** Classical and Neural LMs; Attention/Transformers (BERT/T5/GPT); Chain-of-Though (CoT) Reasoning; Scaling laws; Post-training (instruction tuning, RLHF).
- **Probabilistic Modelling and Reasoning:** Graphical Models; Exact inference (DAGs/HMMs); Causality and Decision-making under uncertainty; Latent variable models; Variational inference and VAEs; Monte Carlo sampling.
- **Programming for Data Science at Scale:** Data-parallel programming model; distributed data-parallel systems; sparse processing; optimization of distributed pipelines; distributed query and graph processing.
- **Current Subjects:** Advanced Topics in ML, Machine Learning Systems and Methods in Causal Inference.

SCHOLASTIC ACHIEVEMENTS

- Recipient of **Kotak-IISc pre-doctoral fellowship** given by Kotak-IISc AIML centre. (website)
- Selected as **Amazon MLSS** mentee, a 4 week program conducted by Amazon's research scientists.

ONLINE CERTIFICATIONS

- **Oxford ML Summer School**: Representation Learning and Generative AI (July 2024)
- **DeepLearning.AI**: Deep Learning Specialization (Feb 2023), Machine Learning Specialization (July 2022)
- **IBM**: Data Science Professional Certificate (Dec 2022) (Code files)
- **Udemy**: Data Structures and Algorithms in Python (July 2022)

SKILLS

- **Programming Languages:** Python (Proficient), C++, MySQL
- **Frameworks:** PyTorch, TensorFlow 2.0, Keras, JAX, FastAPI, Streamlit
- **ML/LLM Tooling:** Hugging Face, vLLM, OpenAI-compatible APIs