Rishi Singhal

Raleigh, NC, USA
rsingha4@ncsu.edu
https://rishi2019194.github.io
linkedin.com/in/rishi-singhal1101

SUMMARY

Passionate researcher dedicated to advance explainable and interpretable AI, with a primary research focus on the dynamics of **memorization** & **generalization** in deep neural networks, spanning generative & classification tasks across NLP & CV. I aim to uncover how models encode these phenomena and how we can mitigate memorization of private & complex data while promoting generalization. My broader research interests include developing methods to improve **efficiency, robustness, safety and privacy** in neural networks.

EDUCATION

North Carolina State University, Raleigh, USA — CGPA: 4.0/4.0
PhD in Computer Science, Advisor: Dr. Jung-Eun Kim — *2025 - 2028 (Expected)*

North Carolina State University, Raleigh, USA — CGPA: 4.0/4.0
Masters in Computer Science, Advisor: Dr. Jung-Eun Kim — *2023 - 2025*

Indraprastha Institute of Information Technology (IIIT) Delhi, India — CGPA: 9.0/10.0
BTech in Electronics and Communication Engg, Advisor: Dr. Rajiv Ratn Shah — *2019 - 2023*

SKILLS AND TOOLS

Python, C++, SQL, MATLAB, Pytorch, Tensorflow, Keras, Scikit-Learn, Numpy, Pandas, SpaCy, NLTK, Nvidia-Triton, MCP, Docker, Flask, Postman, Git

PUBLICATIONS (PUBLISHED AND UNDER REVIEW)

- "Analysing impact of Layer Normalization on Memorization and Generalization", ***Submitted to: NeurIPS (NIPS) 2025.***
- "Distinguishing between Memorization and Generalization at the Feature Level", ***Submitted to: NeurIPS (NIPS) 2025.***
- **Rishi Singhal**, Samyak Jain, Sriram Krishna, Yaman Kumar Singla, Rajiv Ratn Shah, "Beyond Words: A Topological Exploration of Coherence in Text Documents", ***Accepted in: The Second Tiny Papers Track at ICLR 2024.***

EXPERIENCE

Dr. Jung-Eun Kim Lab — Raleigh, USA
Graduate Research Assistant, Mentor: Dr. Jung-Eun Kim — *2024.01–Present*

- Discovered a novel role of LayerNorm (LN) in shaping memorization vs. generalization across Pre-LN and Post-LN transformer models; verified on both **generative** and **classification** NLP & CV tasks, where pruning mere **0.1–0.2%** of total model parameters in Post-LN mitigates memorization by ∼**70%** without degrading generalization, while in Pre-LN it just disrupts generalization.
- Demonstrated that **early LNs** exert the strongest influence on both memorization and generalization in Pre-LN and Post-LN transformers, establishing their critical role in comparison to later layers LNs.
- **Ongoing work:** Theorizing and empirically distinguishing a novel separation of memorization and generalization at the feature level, and studying the impact of residual connections on memorization & generalization in large-scale LLMs (e.g., GPT, LLaMA).

Fermilab — Batavia, USA
Machine Learning Intern, Mentor: Dr. Giuseppe Cerati — *2024.05–2024.08*

- Deployed Graph Neural Networks (NuGraph2/3) on Fermilab's EAF via **Nvidia Triton** & **Docker**, enabling real-time background filtering and semantic labeling for the MicroBooNE neutrino experiment.
- Integrated **Python/C++** client with LarSoft to stream detector data directly to inference servers, reducing memory overhead by **20%** and eliminating intermediate h5 file storage.
- Extended the NuSonic Triton framework to improve scalability and maintainability, with thorough documentation to support DUNE and future Fermilab experiments.
- Contributed production-level code now adopted into Fermilab's official reconstruction pipeline, where it actively supports real-time inference for physics data processing for **10k+ neutrino classification data points**.

MIDAS Lab, IIITD — New Delhi, India
Undergraduate Research Assistant, Mentor: Dr. Rajiv Ratn Shah — *2022.01-2023.04*

- Investigated document coherence as a core metric for evaluating text quality, focusing on its role in downstream NLP tasks such as summarization, machine translation, and question answering.
- Developed Python scripts to apply Topological Data Analysis (TDA) on attention graphs of text documents computed using BERT, RoBERTa, to capture structural and organizational patterns critical for modeling coherence.
- Developed a lightweight MLP-based architecture leveraging TDA features, achieving state-of-the-art performance on the GCDC dataset and outperforming existing transformer baselines by **5%** in accuracy scores.

PROJECTS

Exploring & Analyzing Internal Structure of Language Models to Mitigate Social Biases — GitHub Link

- Investigated the encoding of social biases (e.g., gender, race, sexual orientation) within Pre-trained Language Models (PLMs) like BERT, RoBERTa. Analyzed how, when, and where these biases are embedded in the model's internal layers and neurons, revealing their presence in later model layers and the Feed-Forward Network instead of Attention Heads. This research aims to inform better mitigation strategies like model pruning without relying solely on costly dataset curation or fine-tuning.

Few Informative Data Samples are Good Enough: Introducing Intelligent Data Pruning — GitHub Link

- Developed an 'intelligent data pruning' methodology for under-sampling class-imbalanced datasets, improving model efficiency and accuracy. Outperformed traditional and advanced sampling methods like SMOTE, Gaussian Copula, SDV, RRP, across various software engineering datasets and models, while offering practical insights for large-scale datasets with limited resources. *Recieved best paper mentions in the course.*