Nihal Jain

nihal.b.jain@gmail.com • nihaljn.github.io • nihaljn

Education

Carnegie Mellon University (CMU) **Pittsburgh, PA**
MS in Machine Learning, GPA: 4.08/4.0 *Aug 2021 – Dec 2022*

- *Courses*: Probabilistic Graphical Models, Convex Optimization, Deep Reinforcement Learning

Birla Institute of Technology & Science (BITS), Pilani – Hyd. Campus **Hyderabad, India**
BE in Computer Science, GPA: 9.94/10.0 *Aug 2017 – Jul 2021*

- *Courses*: Probability & Statistics, Machine Learning, Information Retrieval, Deep Learning
- *Honors*: **Gold medal** awarded for securing 1st place among graduating students

Selected Publications **(* = equal contribution)**

On Mitigating Code LLM Hallucinations with API Documentation
Nihal Jain, Robert Kwiatkowski, Baishakhi Ray, Murali Krishna Ramanathan, Varun Kumar
International Conference on Software Engineering (ICSE) 2025 – Software Engineering in Practice Track

CrossCodeEval: A Diverse and Multilingual Benchmark for Cross-File Code Completion
Yangruibo Ding, Zijian Wang*, Wasi Ahmad*, Hantian Ding, Ming Tan, Nihal Jain, Murali Krishna Ramanathan, Ramesh Nallapati, Parminder Bhatia, Dan Roth, Bing Xiang*
NeurIPS 2023 – Datasets and Benchmarks Track

ContraCLM: Contrastive Learning For Causal Language Model
Nihal Jain, Dejiao Zhang*, Wasi Uddin Ahmad*, Zijian Wang, Feng Nan, Xiaopeng Li, Ming Tan, Ramesh Nallapati, Baishakhi Ray, Parminder Bhatia, Xiaofei Ma, Bing Xiang*
Deep Learning for Code (DL4C) at International Conference on Learning Representations (ICLR) 2023
Association for Computational Linguistics (ACL) 2023

Multiviz: Towards visualizing and understanding multimodal models
Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, Ruslan Salakhutdinov
International Conference on Learning Representations (ICLR) 2023

Self-supervised Multi-view Disentanglement for Expansion of Visual Collections
Nihal Jain, Praneetha Vaddamanu, Paridhi Maheshwari, Vishwa Vinay, Kuldeep Kulkarni
Web Search and Data Mining (WSDM) 2023

Generating Compositional Color Representations from Text
Paridhi Maheshwari, Nihal Jain, Praneetha Vaddamanu, Dhananjay Raut, Shraiysh Vaishay, Vishwa Vinay
International Conference on Information & Knowledge Management (CIKM) 2021

Inspiration Retrieval for Visual Exploration
Nihal Jain, Praneetha Vaddamanu, Paridhi Maheshwari, Vishwa Vinay, Kuldeep Kulkarni
NeurIPS 2021 – Workshop on Machine Learning for Creativity and Design

A novel approach to use semantic segmentation based deep learning networks to classify multi-temporal SAR data
Aryan Mehra, Nihal Jain, Hari Shanker Srivastava
Geocarto International 2020

Research Experience

Amazon Web Services (AWS) **New York City, NY**
Applied Scientist II *Oct 2024 – Present*

- Co-led the post-training efforts for the **Amazon Q (formerly AWS CodeWhisperer)** Code LLM
- Integrated preference fine-tuning of LLMs with the Direct Preference Optimization (**DPO**) algorithm
- Curated diverse, high-quality synthetic training corpora with ∼**1M+** samples for coding tasks
- Improved accepted characters with Amazon Q by ∼**16%** indicating enhanced developer productivity
- Collaborating with **Amazon AGI** to advance language model agents on software engineering tasks

Applied Scientist I *Feb 2023 – Sep 2024*

- Developed **execution evaluations** to test API invocation and problem solving abilities of LLMs
- Expanded Amazon Q to support Infrastructure as Code (**IaC**), highlighted at **AWS re:Invent 2023**
- Evaluated LLMs like GPT-3.5 on repository-aware coding tasks (**NeurIPS 2023** D/B track)
- Reduced API hallucinations in LLMs like GPT-4o using RAG (submitted to **ICSE 2025** SEIP)

Applied Scientist Intern *May 2022 – Aug 2022*

- Implemented distributed pre-training of LLMs using PyTorch Lightning and Deepspeed
- Improved language models on code search and generation tasks using contrastive learning (**ACL 2023**)

Adobe Research **Bangalore, India (Remote)**
Research Intern *Jan 2021 – Jul 2021*

- Implemented a self-supervised algorithm for disentanglement of image representations in PyTorch
- Enhanced view-specific (color, style, *etc.*) image retrieval for applications like Photoshop (**WSDM 2023**)

Research Intern *May 2020 – Aug 2020*

- Developed GANs to generate color profiles condition on textual queries for image search (**CIKM 2021**)
- Conceptualized an end-to-end deep algorithm for color-based image editing guided by textual input only

Indian Space Research Organisation (ISRO) **Dehradun, India**
Research Intern *May 2019 – Jul 2019*

- Implemented semantic segmentation of Indian terrain satellite imagery (**Geocarto International 2020**)

Miscellaneous Projects

Datahawk 🌐 **Jan 2025 – Feb 2025**

- Developed a Python package for browsing text data to aid NLP and LLM researcher data workflows

Language Model Reasoning in Base64 🌐 **Sep 2024 – Oct 2024**

- Identified ∼**51%** generalization gap in GPT-4o between English and base64 arithmetic performance

Multimodal Prompting for Image Generation 🌐 **Sep 2022 – Oct 2022**

- Conceptualized combining text and image inputs for prompting Stable Diffusion via **CLIP** representations

MultiViz: Visualizing and Understanding Multimodal Models 🌐 **Jan 2022 – May 2022**

- Built an interpretability framework for multimodal models using **LIME** and **gradient analysis**

Brain Meshing 🌐 **Jan 2020 – May 2020**

- Implemented the Marching Cubes algorithm to extract human brain meshes from 3D MRI scans

Awards & Honors

SIGIR Student Travel Grant **Feb 2023**

- Offered a travel grant to present paper at WSDM 2023 in Singapore

Institute Gold Medal – BITS, Pilani **Aug 2021**

- Awarded a Gold Medal for obtaining the first position in a class of ~1100 graduating students

Merit Scholarship – BITS, Pilani **Jul 2021**

- Received full tuition-waivers throughout undergraduate studies, offered to the top 1% of students

KC Mahindra Scholarship for Post Graduate Studies Abroad **Jul 2021**

- Awarded scholarship to pursue post-graduate studies abroad offered to ~50 out of >1000 applicants

Google AI Summer School **Aug 2020**

- One of the ~50 participants out of >1000 applicants selected to attend the Computer Vision track

Academic Services

Paper Reviewer **2023 – Present**

- NeurIPS (2023), ARR (Feb 2024, Jun 2024), ICML (2024), AAAI (2025), ICLR (2025)

Admissions Committee – CMU Machine Learning Department **2021 – 2022**

- Reviewed ~60 applications to the MS in Machine Learning program for the Fall 2022 intake

Teaching

Teaching Assistant – BITS, Pilani **2018 – 2019**

- *CS F111* - Introduction to Computer Programming
- *CHEM F111* - General Chemistry

Invited Talks

Code Generation Hallucinations & Alignment: An Industry Perspective 🌐 **Nov 2024**

- Columbia University, Host: Prof. Baishakhi Ray

References

Prof. Baishakhi Ray ✉ 🌐	Associate Professor, Columbia University
Dr. Murali Krishna Ramanathan ✉ 🌐	Principal Applied Scientist, Amazon Web Services
Dr. Vishwa Vinay ✉ 🌐	Machine Learning Lead, Canva
Prof. Tathagata Ray ✉ 🌐	Professor, BITS Pilani