Manisha Mukherjee

mmukherj@andrew.cmu.edu
manishamukherjee.github.io

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Research Focus

My research focuses on making large language models effective for software engineering by integrating domain knowledge from human expertise and machine feedback.

Education

Carnegie Mellon University — Pittsburgh, PA
Ph.D. in Software Engineering — 2020-2026
Advisor: Dr. Vincent Hellendoorn; Thesis: Beyond Scaling: Integrating Domain Knowledge into LLM Code Generation

The Pennsylvania State University — University Park, PA
M.S. in Computer Science and Engineering
Advisor: Dr. Thomas La Porta; Thesis: Real-time traffic flow estimation under QoI constraints

West Bengal University of Technology — Kolkata, India
B.Tech. in Computer Science and Engineering

Experience

Carnegie Mellon University — Pittsburgh, PA
Graduate Research Assistant — August 2020-Current

- Trained domain-specialized language models (125M and 762M parameters) from scratch on StackOverflow data, outperforming much larger general-purpose models on code labeling tasks at <$2K training cost. Models released publicly on Hugging Face.
- Developed SOSecure, a retrieval-augmented generation system that leverages StackOverflow security discussions to identify and fix vulnerabilities in LLM-generated code, achieving 72-97% fix rates vs. 38-56% for GPT-4 alone across three benchmarks.
- Applied reinforcement learning with compiler and execution feedback to improve the readability and semantic correctness of decompiled code.

Pacific Northwest National Laboratory — Remote
Research Intern — Summer 2025

- Applied reinforcement learning to IBM Granite vision-language models to improve physics and scientific reasoning on PhysBench multimodal QA tasks.

Adobe Research — San Jose, CA
Research Intern — Summer 2024

- Built a knowledge-graph based RAG system for proprietary enterprise documents using entity resolution, deduplication, and confidence-based filtering to construct a low-noise knowledge graph with document provenance.
- Reduced irrelevant LLM responses by >50% and increased answer relevance by 88% compared to production baselines.

Lawrence Livermore National Laboratory — Livermore, CA
Research Intern — Summer 2023,2022

- Built ML models for large-scale HPC telemetry, including error log classification and power net load forecasting.

Fujitsu Labs America Sunnyvale, CA
Research Intern Summer 2021,2019

- Developed retrieval and learning based approaches for semantic code search and code recommendation.
- Analyzed dataset and retrieval characteristics to inform automated pipeline synthesis and model selection in Fujitsu's AutoML systems.

Idaho National Laboratory Idaho Falls, ID
Research Intern, INL Wireless Security Institute Summer 2020

- Developed ML models for wireless signal classification and threat detection, and built a real-time spectrum monitoring and visualization tool.

Cisco Systems, Inc San Jose, CA
Software Engineer, ASR9K group October 2014-August 2017

- Developed distributed router architecture using SDN to disaggregate control and data planes, implementing OpenDaylight plugins and REST APIs for router cluster management.

Scholarships and Awards

Sansom Graduate Fellowship in Computer Science	2024
Presidential Fellowship in SCS	2023
Frank J. Marshall Graduate Fellowship	2018
Carnegie Institute of Technology Dean's Fellowship	2017
Center for Integrated Healthcare Delivery Systems (CIHDS) Scholarship	2012

Teaching

- **Teaching Assistant** at Carnegie Mellon University Fall 2022,2019
  Principles of Software Construction: Objects, Design, and Concurrency (17-214)
  INI MSIT Project Practicum (14-798)
- **Teaching Assistant** at Pennsylvania State University Fall 2013
  Communication Networks (CMPEN 362)

Selected Publications and Patents

[1] **M. Mukherjee** and V. J. Hellendoorn, "Sosecure: Safer code generation with rag and stackoverflow discussions", *arXiv preprint arXiv:2503.13654*, 2025.

[2] **M. Mukherjee**, S. Kim, X. Chen, D. Luo, T. Yu, and T. Mai, "From documents to dialogue: Building kg-rag enhanced ai assistants", *arXiv preprint arXiv:2502.15237*, 2025.

[3] **M. Mukherjee** and V. J. Hellendoorn, "Skill over scale: The case for medium, domain-specific models for se", *Proceedings of the 2025 IEEE/ACM Second International Conference on AI Foundation Models and Software Engineering*, 2024.

[4] V. J. Hellendoorn, J. Tsay, **M. Mukherjee**, and M. Hirzel, "Towards automating code review at scale", in *29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '21)*, 2021.

[5] **M. Mukherjee**, M. Bahrami, and W. P. Chen, "Source code retrieval", in *US Patent Application 17/085,894*, 2020.

[6] **M. Mukherjee**, J. Edwards, H. Kwon, and T. F. La Porta, "Quality of information-aware real-time traffic flow analysis and reporting", in *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, IEEE, 2015, pp. 69–74.

[7] **M. Mukherjee**, "Determination of real-time traffic flow parameters in different devices based on qoi requirements", in *MS Thesis*, 2014.