Sanidhya Vijayvargiya

Github: sani903 Email: sanidhyv@andrew.cmu.edu
+1 412-708-8214 LinkedIn

Education

- **Carnegie Mellon University** GPA: 4.17/4.33
  Masters in NLP and Machine Learning (MIIS, Language Technologies Institute) *Dec 2025 (expected)*
  - Ongoing work: Enhancing **open-source AI SWE agent (OpenHands)** to effectively handle ambiguous prompts, boosting solution accuracy and resource efficiency in software development tasks advised by **Dr. Graham Neubig**. Graduate Teaching Assistant for Advanced NLP
  - Relevant coursework: Advanced NLP, Generative AI, Search Engines, Introduction to Deep Learning, Machine Learning Systems, Speech Technology for Conversational AI

- **Birla Institute of Technology & Science, Pilani** GPA: 9.22/10
  Bachelor of Engineering - Computer Science *July 2024*
  - Published **8 research papers** in NLP applications for code and software engineering, focusing on productivity improvements in development environments. Designed innovative solutions for code refactoring, software requirements classification and malware prediction in code. Awarded **Best Paper** at ICCSA 2022.
  - Investigated **in-context learning** in LLMs by discouraging the formation of induction heads as a continuation of the work done by Anthropic in *In-context learning and induction heads.*
  - Relevant coursework: Natural Language Processing, Data Mining, Reinforcement Learning, Information Retrieval

Experience

- **MITACS Globalink Internship** Dalhousie University, Canada
  Research Intern *June 2023 - April 2024*
  - Led a project using fine-tuned MLMs and **RLHF** to address **identifier renaming in code**, increasing code readability and maintainability.
  - Curated a dataset of 236,745 high-quality variable name–code snippet pairs. Developed a novel **code readability assessment tool** to measure improvements from proposed model.
  - The 125M parameter renaming model outperformed Gemini Pro by 62.5% and the original identifier names by 22%. The model is accessible on HuggingFace.

- **Undergraduate Thesis at Nanyang Technological University** Singapore (Remote)
  Research Intern *July 2023 - April 2024*
  - Enhanced **code-switched text generation** by integrating grammatical information into multi-head fine-tuning of RoBERTa on POS tag prediction and masked language modeling tasks.
  - Applied few-shot prompting with Chain-of-Thought (CoT) reasoning for **multi-lingual LLMs** such as PaLM and SeaLLM, improving text generation quality between code-switch points.
  - Generated augmentation sentences had **48% lower perplexity** than those produced by GPT-3.5 and led to a 4% decrease in perplexity when added to the SEAME dataset.

Selected Publications

- **Enhancing Identifier Naming Through Multi-Mask Fine-tuning of Language Models of Code**: First author for paper published at 24th IEEE International Conference on Source Code Analysis and Manipulation 2024. **Paper**
- **Software Requirements Classification using Deep-learning Approach with Various Hidden Layers**: First author for paper published at 17th Conference On Computer Science and Intelligence Systems FedCSIS 2022. **Paper**
- **Empirical Analysis for Investigating the Effect of Machine Learning Techniques on Malware Prediction**: First author for paper published at 18th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2023). **Paper**

Projects

- **Synthetic data generation for LLM routers**: Trained various **LLM router** architectures solely on synthetically generated preference data between models that did not previously have any training data and outperformed generic routers by 12%.
- **RAG model for domain-specific queries**: Developed a Retrieval-Augmented Generation (**RAG**) pipeline with Llama 3.1, optimizing for domain-specific queries, achieving top performance in the class employing techniques such as cross encoders and HyDE.

Skills

- **Languages**: Python, TensorFlow, Pytorch, C++, MySQL, SQLite
- **ML**: Natural Language Processing, Reinforcement Learning, Information Retrieval, Image Generation, AI agents
- **NLP**: RLHF, Code Generation, Summarization, Sentiment Analysis, Multilingual NLP, Search Engines, PEFT