Siddharth Yayavaram

syayavar@andrew.cmu.edu | siddharth.yayavaram@gmail.com | (412) 689-0376 | Work Authorization: US Citizen

Education

Carnegie Mellon University Dec 2026
Master of Science in NLP/ML, Language Technologies Institute, School of Computer Science Pittsburgh, PA
Current Coursework: Advanced Natural Language Processing, Generative AI, Machine Learning

Birla Institute of Technology and Science, Pilani July 2025
B.E. in Computer Science (**CGPA: 9.97/10**, Institute ***Gold*** Medalist) Pilani, India

Publications [All First/Co-First Author]

CAIRE: Cultural Attribution of Images by Retrieval-Augmented Evaluation. ICCV'25
Critical Evaluation of Generative Models and their impact on Society @ ICCV'25 | Paper

BERT-based Idiom Identification using Language Translation and Word Cohesion. LREC-COLING'24
Multiword Expressions and Universal Dependencies @ LREC-COLING | Paper

Interpretable Feature Optimization for Sadness Recognition in Speech Emotion Analysis. IEEE IS'24
IEEE 12th International Conference on Intelligent Systems (IS) | Paper

Experience

Carnegie Mellon University Pittsburgh, PA
Research Intern (Undergraduate Thesis), NeuLab | Advisor: Prof. Graham Neubig | Code May 2024 – Mar 2025

- Developed a novel metric to quantify cultural relevance of real and generated images, and built an efficient large-scale (6 million entities) text-disambiguation image retrieval system using FAISS, surpassing SOTA LVLMs on the FOCI benchmark.
- Augmented LLMs with retrieved cultural context and Chain-of-thought prompting to compute relevance across cultural proxies, achieving **+28% F1** on a challenging hand-curated validation set. Achieved Pearson $r > 0.65$ vs human annotations on a dataset comprising universal concepts. Accepted @ **ICCV-W** & currently under review @ (**ACL Rolling Review**).

Nanyang Technological University Singapore
Research Intern, SpeechLab | Advisor: Prof. Chng Eng Siong | Code Mar 2024 – Sep 2024

- Fine-tuned LLaMA-3.1-8B with LoRA on the DAIC-WOZ dataset for text-based depression detection, achieving a **+7.1%** F1 improvement over prior work. Designed a PHQ-8–guided prompting strategy, enhancing both accuracy & interpretability.

Amazon, Applied Science Bangalore, India
Summer Intern | Advisor: Abhishek Persad | Code May 2023 – Aug 2023

- Designed outlier detection metrics and regression models for shipping-cost anomalies, built a Django REST API over UPS data to compute benchmark costs, and implemented BERT-based NER to extract product information for KB construction.

BITS Pilani India
Research Assistant | Engaged in **4** Research Projects in Machine Learning–based Systems July 2023 – May 2025

- **BERT-based Idiom Detection:** Designed custom loss functions to improve token-level idiom recognition. | Code
- **Interpretable SER:** Metaheuristic feature selection for emotion detection; SOTA F1 across 4 popular datasets | Code
- **Malware Detection:** GNN/Sequence models for multi-class classification on imbalanced, obfuscated datasets | Code
- **In-Context-Learning with Information Retrieval:** Critically evaluated the methodology of the ECIR Best Paper, identifying flaws and proposing corrections, achieving improved performance on downstream NLP classification tasks.

Projects

⋆ **Basic PASCAL Compiler | Code** Jan 2024 – May 2024
Implemented a simplified Pascal compiler with LEX/YACC: lexer, parser, semantic checks and intermediate code generation.
⋆ **Show Chain (Blockchain ticketing) | Code** Feb 2023 – Apr 2023
Built a blockchain-based movie-ticket distribution system in Java with ZK-proofs for secure transparent transactions.
⋆ **GO MART (E-commerce)** Sep 2022 – Dec 2022
Contributed to a SpringBoot + Vue.js/Tailwind e-commerce app using OOP principles and RESTful design.

Skills

Programming & OS: Python, C/C++, Java, SQL, Linux, High Performance Computing Clusters (HPC)
Libraries and Frameworks: PyTorch, TensorFlow, Numpy, Pandas, Scikit-Learn, HuggingFace, Matplotlib, spaCy
ML: Natural Language Processing, Diffusion Models, Information Retrieval, Computer Vision, Multimodal ML, GNNs