Siddhant Waghjale

swaghjal@andrew.cmu.edu | (412) 390-6437 | LinkedIn: siddhantwaghjale

Education

Carnegie Mellon University, School of Computer Science — Pittsburgh, PA
M.S. in Intelligent Information Systems (Machine Learning and Natural Language Processing), GPA: 4.06/4.0 — *Dec 2024*

- Coursework: Introduction to Machine Learning, Advanced Natural Language Processing, Question Answering, Multimodal Machine Learning, Neural Code-Generation, Visual Learning & Representations

National Institute of Technology, Karnataka — Karnataka, India
Bachelor of Technology in Information Technology, GPA: 8.17/10 — *June 2020*

Experience

Tesla — Palo Alto, USA
Autopilot | Software Intern — *May'24 - Aug'24*

- Developed an **AI-Oncall Chatbot** that resolved **50%** of user queries without human intervention by using advanced **RAG** techniques, real-time vector database updates, and multi-channel integration; and was adopted by over **300** software engineers
- Created a **failure summary** feature for **Teamcity builds** using an LLM agent-based summarizer, reducing build issue resolution time by **80%**.
- Reduced overall build time by **60x** by optimizing the SDK download and extraction process through improved **caching logic**

Kaleyra — Bangalore, India
Chatbot Team | Senior Associate Software Engineer (Team Lead) — *Sept'20 - June'23*

- Led R&D team in developing a scalable **Chatbot** with advanced NLU using Transformers, achieving **350** TPS and serving **10M** customers daily while reducing AWS costs by **$823/month**
- Engineered Email and SMS **Spam Detection** engines by fine-tuning BERT and optimizing with ONNX for a **20x** boost in classification latency and throughput
- Devised an **automatic authentication system** with document and face verification **(AWS ISV Innovation Cup finalist)**

Deloitte Consulting USI — Bangalore, India
Oracle Cloud Solutions | Software Intern — *May'19 - July'19*

- Developed expertise in real-time business processes, **BI reporting**, and data modeling by creating data models, formulating **PL/SQL** queries, and achieving key project milestones

Publications

ECCO: Can We Improve Model-Generated Code Efficiency Without Sacrificing Functional Correctness? (Under Review)
Siddhant Waghjale, Vishruth Veerendranath, Zora Zhiruo Wang, Daniel Fried

Academic Projects

Two-Stage Multimodal Architecture for Visual Abductive Reasoning — *Jan'24 - April'24*

- Developed a **two-stage training architecture**, enhancing image captioning models to infer commonsense information from images by reasoning beyond just identifying objects
- Improved linguistic reasoning in multimodal data by **fine-tuning** vision and language models, achieving notable gains

Improving Performance for Multi-Table Question Answering — *Jan'24 - April'24*

- Implemented two approaches for multi-table QA using **code-based models** and **chain-of-table reasoning**
- Improved evaluation metrics and conducted extensive quantitative and qualitative analysis to enhance performance

Detecting LLM Generated Text in Multi-generator, Multi-domain, and Multi-lingual Black-Box Setting — *Aug'23 - Dec'23*

- **Fine-tuned** Mistral 7B and Llama2 7B on the M4 dataset for monolingual and multilingual text detection, achieving peak accuracies of **92%** and **77%**, respectively
- Performed **paraphrase tests**, showing that the Logistic Regression model generalized better than fine-tuned models

Skills

Languages: Python, Golang, C++
Technologies & Tools: PyTorch, TensorFlow, Langchain, Huggingface, ONNX, CUDA, Kubernetes, Docker
Databases: MongoDB, Redis, SQL, Object Storage, ChromaDB, Snowflake, PostgreSQL, RabbitMQ