Soham Dinesh Tiwari
soham.tiwari800@gmail.com | sohamdtiwari.com | LinkedIn: soham-tiwari | Google Scholar | +1 412-909-7089

EDUCATION

Carnegie Mellon University (CMU) Pittsburgh, PA
Master of Science in Intelligent Information Systems (MIIS) | CGPA 4.0 / 4.0 Dec 2023
Graduate Courses: *Advanced NLP, Multimodal Machine Learning, Speech Processing, Deep Reinforcement Learning*
Graduate Teaching Assistant: Multimodal Machine Learning 11-777 - Spring 2023, Fall 2023, Spring 2024

SKILLS

Programming Languages: Python, Java, C/C++, JavaScript, SQL, Swift, Bash
ML Tools & Frameworks: PyTorch, TensorFlow, HuggingFace, OpenAI, Pandas, Scikit-learn, Matplotlib, Numpy, SciPy, PySyft, PineCone, DGL
MLOps Tools & Frameworks: Wandb, Mlflow, Milvus, ZenML, Flask, AWS, Docker, Kubernetes, Langchain, Streamlit, Dask, OpenSearch

WORK EXPERIENCE

AppZen San Jose, CA
Data Scientist - DS2 Feb 2024 - Current

- Developed **AI** Expense Audit **Fraud Detection** system, helping Airbus, Databricks and more, catch ~$91,000 in fraudulent T&E expenses.
- Fine tuning LLMs like **Llama-2, Llama -3.2** using **LoRA adapters** and trained **Graph Convolution Networks** and **Naive-Bayes models**.
- The novel fraud detection system improved the recall of the previous Decision Tree based duplicate system by more than 50%.
- **Designed** and scaled up **ETL** architecture - Dask, Kubernetes and OpenSearch Index **reducing computation time** by more than **200%**.
- **Curated** multimodal fraud **evaluation dataset**, to **benchmark** different models, and created Llama-based **agent** to detect suspicious users.
- Gathered post-implementation feedback from customers, and worked with the CSM team to maximize value to customers.

Apple Seattle, WA
AIML Natural Language Intern | Siri NL Input Representations May 2023 - Aug 2023

- Developed framework for **LLM** distillation/optimization using Jax, supporting **distributed GPU training**, **tensor** and **model sharding**.
- **Mitigated** loss convergence issues in LLM **distillation** using **TinyBERT** to put **large language models on-device** using developed tool.
- Increased **BLEU** scores by **2%** and improved LM **perplexity**, **bits-per-byte** by **47%**. Contributed to LLM LoRA adapters, PEFT library.

Nanyang Technological University Singapore, Singapore | Remote
Research Intern Aug 2021 - Jul 2022

- Improved sound event detection **F1 scores** by **3%** on AudioSet dataset and DCASE challenge dataset.
- Enhanced **log-Mel spectrograms** and **deep pre-trained audio neural network** (PANN) using frequency dynamic **convolutions**.
- Developed **novel curriculum-learning** algorithm increasing **SPIDEr** scores by **5%** for Automated Audio Captioning system.
- Accepted paper in APSIPA 2022. Published First-author poster at **NeurIPS** 2021 and paper in **IJACSA** on related work.

Forty4Hz INSIA Bangalore, India | Remote
React and Data Science Intern Jan 2022 - Jun 2022, Sep 2020 - Jun 2021

- Developed React/Node.js search platform to visualize **business intelligence insights** with interactive charts, following Agile practices.
- Built large-scale python **data extraction, processing** & **ingestion** engine for Excel and CSV files, from client **data warehouses**.
- Deployed production **RNNs**, **Bi-LSTM** for **modeling time-series data** and generating insights both in real time and offline reports.

Gravitas AI London, UK | Remote
Natural Language Processing Engineering Intern Aug 2021 - Oct 2021

- Used **Graph Neural Networks** and **BERT transformer** for NER, co-reference resolution, relationship extraction, **NLU/NLP** tasks.
- Updated ChatBot's Neo4j **knowledge graph** & **medical ontology** by using NLP text processing pipeline to parse **terabytes** of texts.

University of British Columbia Vancouver, Canada | Remote
MITACS Globalink Research Intern May 2021 - Aug 2021

- Researched and decoded **EEG** signals using DL, ML algorithms like **SVMs**, **K-Means clustering**, Linear Regression & **signal processing**.
- Designed experiments to hypothesize that infants can distinguish living and nonliving things, while preserving privacy of the data.
- Verified significance of experimental results using **grid search**, and **statistical** tests like **p-tests** and **t-tests**.

PROJECTS

Chart generation from NL queries Jan 2025

- Developed an AI agent, detects which information the team is from and surfaces relevant charts that have been changed recently.
- Created python package from ground-up supporting cross-LLM compatibility - AWS SageMaker, OpenAI, DeepSeek, HuggingFace.
- The Agent can also generates Vega-Lite charts to answer users' NL queries about recent data changes.

End-to-end training from scratch of GPT-2 LLM and PaLI-Gemma VLMs Nov 2024

- Trained LLM, VLM from scratch, implementing Multi-Head Attention, Grouped Query Attention, Rotary Positional Embeddings (RoPE).
- Reduced GPT-2 inference time from 1000ms to 93ms through mixed precision training (FP16/BF16), gradient scaling, and torch.compile.
- Used DistributedDataParallel pipeline with AdamW, gradient accumulation, and cosine decay scheduling.
- Used CLIP/SigLip contrastive learning for vision-language alignment. Trained the LLM using FineWeb-Edu dataset.
- Evaluated system using HellaSwag benchmark, with efficient KV-caching and sampling strategies (Top-P, temperature control).

End-to-end production customer satisfaction prediction using MLOps Dec 2023

- **Improved** customer product satisfaction regression R2 score by **12%** applying ML algorithms like LightGBM, XGBoost, RandomForests.

- Conducted hyperparameter optimization with Optuna, monitored training with MLflow and Wandb for best hyperparameter identification.
- Implemented data ingestion, processing, train-test-split steps, followed by automatic model training & evaluation using RMSE, R2 scores.
- Enabled CI/CD support with automatic model inference API deployment using MLflow and Docker using model performance triggers.

On Device Generative AI Interactive Multimodal Mock Interview Conversational Agent | CMU Aug 2023 - Dec 2023

- Built low latency agent that runs **on-device**. Curated dataset, built **multimodal perception** model for facial emotions image classification.
- Used **retrieval augmented generation** (RAG) with Langchain, Milvus Vector DB search and **ChatGPT/GPT-4/Llama-2** for code hints.
- Used **prompt engineering**, **Chain Of Thought** and few-shot prompting to ensure generative AIs don't leak direct solutions in responses.
- Built **real-time**, **multithreaded multimodal** processing backend using PSI, NodeJS, FastAPI. Prototype accepted at ISLS 2023.

Embodied Vision and Language Navigation | CMU Aug 2022 - Dec 2022

- Improved agent's collision recovery ability in the Room-Across-Room dataset's simulated homes using vision, depth and text modalities.
- Implemented self-orienting heuristics, **PPO**, **cross-modal attention**, and **multimodal** alignment-based RL reward functions.
- Used **LSTMs, ResNets** and **VLN-BERT** architectures to implement text-guided waypoint prediction strategy, improving **SPL** by **1.9%**.

Language-Agnostism, ChatGPT (LLM) Query Rewriting for Multilingual Document QA| CMU Jan 2023 - May 2023

- Increased **Recall**@1 by **22%** over XLM-RoBERTa retriever-reader using **Dense Passage Retrieval** with **LaBSE Sentence** Transformer.
- Identified limitations of LLMs in zero-shot **query rewriting** and **dialog summarization** for document grounded question answering
- Used Fusion-in-Decoder with mT5 and mBART to improve **multilingual** text **generation**. Accepted paper at **ACL** DialDoc 2023.

Cascaded Code-switched Speech to Monolingual Speech Translation | CMU Jan 2023 - May 2023

- Developed **cascaded**, speech to speech machine translation system for **code-switched** Indic languages - English, Bengali, Sanskrit.
- Used CNNs, ConvNeXt and transformers like **Branchformer**, **Conformer** for speech to text translation, increasing **BLEU** by **9.6 points**.
- Used state of the art text to speech synthesis models like Tacotron2, FastPitch, HiFiGAN for the Prabhupadavani dataset..

Optimal Active Learning for Efficient Multilingual Fine-tuning | CMU Aug 2022 - Dec 2022

- Devised **active learning** algorithm using **KNNs, mBERT embeddings,** sample **uncertainty** to find informative data points for fine-tuning.
- **Reduced data labeling** effort by **80%**. Conducted experiments with Low-Resource and High-Resource Languages from XNLI & MARC.

SELECTED PUBLICATIONS

Srinivas Gowriraj*, **Soham Dinesh Tiwari***, Mitali Potnis, Srijan Bansal, Teruko Mitamura, Eric Nyberg. "Language-Agnostic Transformers and Assessing ChatGPT-Based Query Rewriting for Multilingual Document-Grounded QA." ACL 2023

Vitiello, Rosanna, **Soham Dinesh Tiwari**, R. Charles Murray, and Carolyn Rosé. "Traveling Bazaar: Portable Support for Face-to-Face Collaboration." ISLS 2023

Koh, Andrew, **Soham Dinesh Tiwari**, and Chng Eng Siong. "Automated Audio Captioning with Epochal Difficult Captions for curriculum learning." APSIPA 2022