I am working on developing multi-modal foundation models for specialized fields such as healthcare and biology. Within that I specifically developed methods in the following areas:

1) **Synthetic data generation with multi-level optimization-based feedback mechanism:** Generating high-quality synthetic data is crucial for addressing data scarcity in specialized domains. To tackle this challenge, we employ language models (LMs) to generate task-specific synthetic data. Our approach is grounded in two key strategies: (1) fine-tuning LMs on an abundant, related dataset to produce data tailored to the downstream task, and (2) introducing a novel multi-level optimization framework to further ensure that synthetic data generation is highly relevant to the downstream task. The multi-level optimization framework dynamically learns instance-specific weights to samples in the abundant source dataset, prioritizing those most relevant to the target task. These weighted examples are then used to fine-tune the LM, effectively bridging the gap between the source and target data. This framework ensures that the generated synthetic data aligns closely with the target task requirements, thereby maximizing its utility. Our method has shown significant improvements in performance. On NLP benchmark datasets, it achieved an average of **10% increase in accuracy** over other augmentation baseline methods in low-data settings. The results were published in Transactions of the Association for Computational Linguistics (TACL) [1]. Furthermore, we applied this approach to classify long COVID-related articles, an area with scarce annotated data, and demonstrated a **13% improvement in accuracy** compared to traditional keyword-based search methods. Our work is published in Nature Scientific Reports [2].

2) **Continual learning in sequential data scenarios:** Our work in continual learning focuses on developing an approach based on meta-learning that can identify an optimal subnetwork within a large language model (LLM) for fine-tuning on a specific downstream dataset. By fine-tuning only this subnetwork and keeping other weights fixed to pre-trained values, we achieve enhanced performance over vanilla fine-tuning and parameter efficient fine-tuning (PEFT) methods, improving performance across different LLMs. This work was presented at the North American Chapter of the Association for Computational Linguistics 2024 [3]. For continual learning, we extend this framework to dynamically adapt a small subnetwork to incoming sequential data with a large part of the network fixed to the pre-trained weights, enabling us to improve performance on current tasks while maintaining performance stability across previously seen data. This

approach has proven effective in our experiments on early sepsis prediction across a sequential ordering of hospital datasets, demonstrating the practical impact of sequential data adaptation and model reusability. We are preparing our manuscript for publication at Nature Methods.

3) **Security of deployed models:** The widespread adoption of large language models (LLMs) has brought attention to the challenge of distinguishing human-authored texts from machine-generated ones. This distinction is critical to mitigating risks associated with misinformation and safeguarding against representation collapse—a phenomenon that arises when an over-reliance on machine-generated data for training erodes the diversity and quality of future models. Statistical watermarking offers a promising solution for reliably detecting LLM-generated texts. These techniques embed hidden patterns in LLM-generated texts that, while imperceptible to humans, can be detected by algorithms. However, existing methods often degrade text quality after watermarking. We propose a robust watermarking technique for LLMs that embeds statistical watermarks into generated texts while preserving their quality. Our approach is grounded in a multi-objective optimization framework that improves detectability while maintaining semantic coherence. Experiments on multiple benchmarks demonstrate that our method achieves an improved Pareto frontier compared to other baseline watermarking methods. Further, we also demonstrated the robustness of our proposed method against various attacks such as paraphrase, copy-paste and corruption attacks. This work was presented at the International Conference on Machine Learning (ICML), 2024 [4].

Foundation model initiatives:

I am leading a research project tackling a key challenge in genomics, particularly within Sequential Fluorescent In Situ Hybridization (seqFISH), which enables single-cell resolution mapping of gene expression. While seqFISH offers advanced mapping capabilities, it frequently faces limitations due to missing data in three-dimensional gene coordinate mapping. To address this issue, we have developed a foundation model that predicts missing gene coordinates using known coordinates as input. The results indicate significant improvements over conventional methods for 3-D coordinate prediction, showcasing the effectiveness of their proposed approach.

1. **Sai Ashish Somayajula**, Linfeng Song, Pengtao Xie. A Multi-Level Optimization Framework for End-to-End Text Augmentation. Transactions of the Association for Computational Linguistics, 2022.

2. **Sai Ashish Somayajula**, Onkar Litake, Youwei Liang, Ramtin Hosseini, Shamim Nemati, David O. Wilson, Robert N. Weinreb, Atul Malhotra, Pengtao Xie. Improving Long COVID-Related Text Classification: A Novel End-to-End Domain-Adaptive Paraphrasing Framework. Scientific Reports. Nature Portfolio, 2024.

3. **Sai Ashish Somayajula**, Youwei Liang, Li Zhang, Abhishek Singh, Pengtao Xie. Generalizable and Stable Finetuning of Pretrained Language Models on Low-Resource Texts. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2024.

4. Mingjia Huo*, **Sai Ashish Somayajula***, Youwei Liang, Ruisi Zhang, Farinaz Koushanfar, Pengtao Xie. Token-Specific Watermarking with Enhanced Detectability and Semantic Coherence for Large Language Models. *Equal contribution. International Conference on Machine Learning, 2024.