"Sometimes innovation is only old ideas reappearing in new guises . . . [b]ut the new costumes are better made, of better materials, as well as more becoming: so research is not so much going round in circles as ascending a spiral."[1]

In the ever-evolving landscape of artificial intelligence, I am captivated by the notion that innovation often breathes life into age-old ideas, donning them in new guises that are not only more aesthetically pleasing but also remarkably robust. My venture into intelligent systems has been a purposeful exploration of this concept. Over time, I have come to realize that true progress lies not merely in chasing novel concepts but in harnessing the potential embedded in the foundations of our knowledge. In one experience, during a self-directed project, I used ideas from statistics and machine learning to create a novel consumer targeting system by modeling the rich data from Augmented Reality systems. I worked with multiple tools and methods, including statistical modeling (to identify augmented visuals influencing the consumer purchase), an exemplar part-based 2D-3D alignment method to find the best matching 3D models of furniture present in the user's preferred purchase viewpoint in an AR app and used a combination of 3D style compatibility and color compatibility algorithms to recommend relevant furniture. Having published this work at ICCVW 2019, I learned how to use ideas from data science and ML to develop useful solutions for systems such as AR. The realization that these 'ideas' can be refined and fortified with modern techniques and materials resonates deeply with my academic and professional pursuits.

During my bachelors at IIT Kharagpur, I undertook several internships in an effort to practically apply the theoretical knowledge I was gaining. As a remote research collaborator with the Sustainability and AI Lab at Stanford University, I helped build and validate machine learning models with remote sensing data for socioeconomic, computational sustainability, and computer vision tasks using tools from object detection, reinforcement learning, noise contrastive estimation, transfer learning, and statistics. The experience of working on these projects kindled my interest in research and was one of the main factors that motivated me to learn math, statistics, computing, and machine learning further.

I also worked at qure.ai, a healthcare startup, as a research intern, where I developed an end-to-end pipeline for classifying and segmenting medical entities (tubes and catheters) in chest X-rays. I devised a two-step strategy for precise tube tip localization using deep learning-based segmentation and advanced image processing techniques. By conducting extensive experiments with various CNN architectures, I surpassed baseline methods and contributed to the successful integration of these models into qure.ai's deep learning stack. This work led to the development of the qXR-BT product, which later received FDA approval for clinical use.

After graduation, I joined Microsoft Turing Research as a Research Fellow, where I got the opportunity to dabble with large-scale models in language (and vision). Having developed an avid interest in language understanding, I subsequently forayed into this area and personally undertook several initiatives that were aimed at developing evaluation frameworks for various LLMs. In one of the projects, I actively sought inspiration from historical ethical frameworks and moral psychology and developed a psychometric assessment tool to measure the ethical reasoning capabilities of LLMs based on Kohlberg's Cognitive Moral Development model and Defining Issues Test. Additionally, I established a comprehensive framework to facilitate the infusion of ethical policies for moral alignment in LLMs by leveraging in-context learning to address complex social dilemmas characterized by conflicting values (value pluralism). This project made me recognize that understanding the ethical implications of LLMs requires a nuanced comprehension of human behavior and moral understanding. It provided valuable insights into the development of ethical frameworks, principles, guidelines, methodologies, and tools essential for the responsible and ethical design, evaluation, and deployment of LLMs. I learned a lot from these projects as I overcame initial failed attempts by iteratively improving for 8 months until successfully publishing as a joint first-author in top-tier avenues like EMNLP, EACL, Neurips Workshop, and LREC-COLING.

In another project, I performed a literary survey of state-of-the-art methods in visual language understanding and co-led the development of DUBLIN, a large-scale transformer-based encoder-decoder model adept at interpreting diverse document types. I successfully trained DUBLIN to excel in a wide array of tasks by using techniques like multitask pre-training, curriculum learning, template-based multimodal instruction tuning, and a novel and optimized method for variable input resolution training. I also realized the importance of thorough data analysis and visualization in gaining valuable intuitions about the problem, allowing me to come up with novel solutions. From implementing basic methods in statistics and machine learning from scratch during undergraduate school to developing these complex systems, this project served as a crucible for the development of my technical acumen. It provided me with a profound understanding of the intricacies involved in crafting large-scale machine learning systems and further catalyzed my interest in weaving together existing methods with modern technological advancements. This work also resulted in multiple patents and a publication in EMNLP 2023, where I was the joint first-author.

I also worked on multilingual language modeling, co-leading the development of sPhinX, a 6.7B parameter language model proficient in 51 languages. This was achieved through instruction tuning on a newly created 1.8M multilingual instruction

[1] Karen Sparck Jones. 1994. Natural Language Processing: A Historical Review, pages 3–16. Springer Netherlands, Dordrecht.

dataset derived from selective translations of Chain-of-Thought (CoT) prompts. I devised a novel, sample-efficient training mechanism by prepending N-shot samples from the same language and task to the training prompts. This approach improved the multilingual capabilities of Phi and Mistral-based models by an average of 10% over state-of-the-art performance on multilingual benchmarks, even without pre-training in 51 languages. The dataset and training mechanism enhanced multilingual performance in production models at Microsoft, including Copilot and M365. This work is currently under review at NAACL 2025.

While the initial waves of these Large Models (including but not limited to LLMs) will be based on the inferencing capabilities and availability of cloud computing, I believe we should prepare for a future wave of "on-device large models." If done right and made to fit into the constraints of such devices, this will not only offer obvious "edge computing" advantages such as improved latency and offline availability but also democratize the reach of large models to emerging markets. In an effort towards this direction, I also explored methods from matrix compression, quantization, early exit decoding, and uncertainty quantification to make these models amenable to single GPU deployment and faster inference. However, such "on-device" prospects also raise concerns about the security and privacy implications of running LLMs on mobile devices. In doing so, I also on several security and privacy-preserving techniques that can be implemented in the design of on-device LLM-based features. One of the directions we explored was based on differentiating instructions from data by employing cryptographic delimiters to prevent man-in-the-middle threats.

There is a significant degree of uncertainty surrounding the use cases, benefits, and potential risks associated with these models. These Large Models represent a new and rapidly evolving field, with more unknowns than "knowns." Over time, new insights, applications, and risks will emerge. I now desire to pursue further research in this area and apply tools from applied math, statistics, ethics, and machine learning towards the ethical development of socioculturally informed, reliable, efficient, and secure ML systems, focusing on large language (and vision) models.

Currently at Harvard as part of the Data and Knowledge Exploration (DtaK) Lab, my research focuses on controlling LLM behavior through techniques such as model editing and representation steering, aiming to mitigate biases without explicit retraining. Inspired by the work "Locating and Editing Factual Associations in GPT," we are investigating whether LLMs like Llama, Gemma, and Phi implicitly store biases as factual associations. In another study, we identified patterns (e.g., a bent circle in three-dimensional space) representing certain concepts in higher-dimensional spaces. This research aims to enhance our understanding of complex concepts and deepen knowledge of LLM internals.

In another project at Harvard, I co-led the development of a course planner leveraging the Harvard Course Database to recommend courses based on students' short- and long-term goals. We implemented a Retrieval-Augmented Generation (RAG) system using ChromaDB's vector database and fine-tuned the LLama 3.1 8B model to deliver responses in a conversational tone. The system simplifies course selection for Harvard students by aligning their academic schedules with career objectives and helping them efficiently plan weekly schedules.