My research involves the study of mechanisms to enhance *understanding* and *usability* of current day language models. On the usability side, my work as a pre-doctoral researcher at Google Research has focused towards modular (and non-monolithic) language models that are lightweight and easy to adapt for downstream use-cases (knowledge, domains). I am excited about **furthering modularity to improve efficiency and control** of current models. Today, a generic large language model (LLM; ∼10B parameters) equates to ∼80 gigabytes of storage—infeasible to store on mobile. Given large models with numerous skills, I am interested in ways to cheaply discover sub-modules that satisfy task and computational requirements: a translation module from an LLM, a sticker generation module from an image generation model. On the understanding end, I am deeply interested in **developing robust tools for identifying inductive biases** in models and utilizing them for enhanced evaluation, control, and modeling. I delved into aligned problems during my bachelor's thesis work at the Technion Israel and during long-term research collaborations at LTI CMU and at CILVR Lab NYU.

Modularity for Usability and Control of Language Models

Overview. Today, large-scale pre-trained models, especially LLMs, are being widely used to retrieve knowledge, curate content, seek clinical advice, among many other applications. However, their monolithic structure poses a number of challenges that limit their usability. Namely, such monolithic models are challenging to control, update with new knowledge, and serve downstream. Consider a conversational system designed to clinically answer maternal queries from rural India. In order to be useful, the system needs to be oriented for reliable clinical knowledge specialized for maternal care and be inexpensive to serve and use. Developing current day systems for these axes is essential for usability yet a difficult problem given our current tools. These problems are even more complex when downstream users or organizations need to perform such adaptation over black-box models on proprietary data. I am interested in designing algorithms to extend and serve these models to make them more functional and useful for real-world problems.

Composition of Language Models. Whilst at Google Research, advised by Dr. Partha Talukdar and Dr. Prateek Jain, I worked on a strata of this usability problem: Efficiently extending general capabilities available through large models to specialized downstream knowledge. We first observed that conventional fine-tuning (training all or a subset of model parameters) for a certain capability quickly leads to regression of other abilities, hence lacking robustness. Instead, we approached the problem of adaptation from a compositional standpoint: a large language model (an *anchor*) can be bootstrapped with new niche knowledge through composition with other, potentially smaller, specialized (*augmenting*) models. We observed that general purpose abilities of an anchor LLM can be extended for settings such as extremely underrepresented languages and code through composition with associated specialized LMs. In our recent **ICLR'24** submission (Bansal et al., 2024), we ground our framework to learn a few parameters over layer representations of the models, keeping original models unchanged, hence also preserving their capabilities.

Given the emphasis on foundational design principles of modularity, model re-use and efficiency in our work, we are now closely working with Google DeepMind and product teams at Google. We are exploring a new realm of serving foundation models to downstream users that allows them to adapt a model to their custom data in a compositional manner. I see this work as an initiation towards a line of work for adaptive, controllable, and hence more useful models.

Going forward, I wish to delve closer into the problem of efficiency—another key pillar towards more useful models. While our current work enables new capabilities through composition of multiple models, I am keen to explore the converse problem: Given a large general model, can we extract multiple specialized models that fit a specified computational budget or task requirements? I believe that such methodologies can utilize sparsity of current models to obtain more practical and deployable versions. This is especially useful in light of recent results where task-specific small models are shown to outperform large generic models (Raffel et al., 2023). Initial lines of work towards this end propose an inherently modified pre-training scheme (Kudugunta et al., 2023) as well as post-hoc algorithms (Zhang et al., 2022), both of which are exciting directions.

Model Understanding and Analysis for Evaluation and Modeling

Overview. As a step towards larger goals, it is essential to enhance our understanding of current systems. I believe that interpretability and analysis of models can provide a systematic way to understand limitations and strengths of current models. Breaking down a model's decision process into a human comprehensible set of inductive biases could open avenues to attribute computation rules behind model predictions, discern how a set of training examples are generalized to learn decision rules, and finally control undesirable behavior. I am especially excited about using robust model understanding tools for use cases in evaluation and modeling.

Evaluating Explanations. *What we cannot measure we cannot improve.* Although a large number of efforts have made strides towards the overarching goal of model understanding, no proper means for evaluation (and hence measuring progress) exist. As a fundamental step towards enhancing model understanding, I joined Prof. Danish Pruthi at LTI CMU to ground and quantify the notion of explanation quality. Our formulation builds from the core use of explanations: A means of communication in any teacher-student setup. Let's say a student A is given some pairs of quadratic equations with their roots, while another student B is additionally given an explanation for how each equation was solved. The relative ability of the two students to solve quadratic equations in the wild is reflective of the quality of explanations given to student B. Our **TACL** paper (Pruthi et al., 2022) formalizes this methodology to evaluate saliency attribution methods for text models.

Evaluating Generalization via Model Analysis. I have utilized model analysis tools to evaluate generalization attributes like reliance on spurious correlations and memorization of individual training examples. During my bachelor's thesis with Prof. Yonatan Belinkov at the Technion, I studied information distribution across neuron activations and discovered strong correlation patterns with distinct generalization behaviors. Our work, presented at **NeurIPS'22** (Bansal et al., 2022), bypasses the necessity for specialized out-of-distribution evaluation sets and introduces intrinsic information-theoretic metrics such as mutual information and entropy as a way to perform model selection.
Similarly, my work with Dr. Naomi Saphra and Prof. Kyunghyun Cho at CILVR Lab NYU studied mode connectivity in the loss surface of real-world text models. For the first time, we discovered that distinct modes (as basins in the loss surface) exist in the loss surface, each of them containing set of models that depict a particular generalization behavior. Our work at **ICLR'23** (Juneja et al., 2023) challenged conventional beliefs of how individual models are connected in the loss surface, paving way to re-think several foundational aspects in weight averaging, model selection, and initialization.

Going forward, I wish to continue to utilize robust model analysis tools for foundational challenges in evaluation and modeling. As an example, I am curious to investigate the role of training dynamics (or, transition patterns in the loss surface) to make optimal design choices for hyperparameters and dataset mixtures for large-scale training. Moreover, I wish to work towards the ambitious goal of identifying intermediate computations behind model outputs. Popular methods in language models like chain-of-thought style reasoning are often misinterpreted as explanations, however past work (Lanham et al., 2023) has established that these generations are often unfaithful to a model's output decision process. I believe that inductive biases grounded in intermediate representations can lead to more trustworthy model understanding, and hence enhanced use for debugging and control.

Why Harvard?

My impetus for pursuing a doctorate program is greatly shaped by my past experiences working closely with my advisors. At Harvard, I am particularly excited to work with **Prof. Martin Wattenberg**. Prof. Watternberg's extensive work on visualization could pave very interesting directions for model analysis and interpretability. I strongly believe that I can further my work and interest in studying loss surface mode connectivity under their guidance, especially given the influence of efficient visualization in studying loss surfaces. I am also genuinely interested in working with **Prof. David Alvarez-Melis** to use insights from model analysis for more interpretable, robust, and controllable models. I am particularly inspired by their past work on gradient flows to interpret datasets and training dynamics. I would also be keen on interacting with **Prof. Boaz Barak** and **Prof. Sham M. Kakade** to further our theoretical understanding of language models and associated scaling laws. I am excited to be a part of the **Kempner Insitute** and closely collaborate with the research fellows at the institute.

References

[1] Rachit Bansal, Bidisha Samanta, Siddharth Dalmia, Nitish Gupta, Shikhar Vashishth, Sriram Ganapathy, Abhishek Bapna, Prateek Jain, Partha Talukdar. LLM Augmented LLMs: Expanding Capabilities through Composition. *International Conference on Learning Representations (ICLR; under review)*, 2024.

[2] Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, Naomi Saphra. Linear Connectivity Reveals Generalization Strategies. *International Conference on Learning Representations (ICLR)*, 2023.

[3] Rachit Bansal, Danish Pruthi, Yonatan Belinkov. Measures of Information Reflect Memorization Patterns. *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[4] Danish Pruthi, Rachit Bansal, Bhuvan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C. Lipton, Graham Neubig, William W. Cohen. Evaluating Explanations: How much do explanations from the teacher aid students?. *Transactions of the Association for Computational Linguistics (TACL)*, 2021.

[5] Devvrit, Sneha Kudugunta, Aditya Kusupati, Tim Dettmers, Kaifeng Chen, Inderjit Dhillon, Yulia Tsvetkov, Hannaneh Hajishirzi, Sham Kakade, Ali Farhadi, Prateek Jain. MatFormer: Nested Transformer for Elastic Inference. *arXiv preprint arXiv:2310.07707*, 2023.

[6] Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun and Jie Zhou. Transformer Feed-forward Layers are Mixtures of Experts. *Association for Computational Linguistics (ACL)*, 2022.

[7] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, et al. Measuring Faithfulness in Chain-of-Thought Reasoning. *arXiv preprint arXiv:2307.13702*, 2023.