Harshvardhan Agarwal Email : hvag976@stanford.edu

Statement of Purpose

of Harshvardhan Agarwal (Ph.D. Applicant)

AI systems have made remarkable progress in sequence-based tasks, yet we still understand surprisingly little about how transformers internally represent structure or how their architectural biases limit generalization. Even core design choices such as the quadratic computation in self-attention impose assumptions about fully pairwise interactions that may not be necessary for many structured domains. Recent linear-time models based on state space formulations have begun to challenge this assumption by showing that strong sequence modeling is possible without dense all-to-all interactions. These developments raise a broader question about how architectural constraints should reflect the structure of the domain, particularly in graphs, relational data, or program-like settings where the information layout is fundamentally different from a sequence. I am broadly interested in developing a deeper understanding of these models, how they form internal representations, what inductive biases they rely on, and why certain reasoning behaviors emerge. My goal is to use these insights to design architectures that can generalize beyond sequences, where structure plays a central role. In particular, I hope to explore how principled modifications, grounded in an understanding of model internals, can lead to more efficient, more interpretable, and more general-purpose reasoning systems.

This motivation first shaped my research during my senior year at IIT Bombay, I worked with Prof. Sunita Sarawagi on analyzing the mechanisms underlying in-context learning for sequence-to-sequence tasks. Prior work showed that shallow transformers can form induction circuits for classification, but we hypothesized that ICL for sequences requires two distinct components: (1) learning the alignment between input and output phrases, and (2) predicting the next tokens given previous. Using a synthetic generator based on probabilistic grammars and classical alignment models, we showed that pretrained LLMs reliably form induction heads for next-token prediction but fail to learn these alignments in-context once sequences grow longer. To address this gap, we developed ICA-Tune, a focused adaptation method that fine-tunes only a small subset of attention parameters, yielding large improvements in accuracy and out-of-distribution generalization. This work [1], published at the *International Conference on Machine Learning*, gave me a clearer mechanistic view of the computations transformers perform internally and strengthened my interest in models that can acquire new structure more effectively.

While pretrained transformers readily adapt to new sequence-modeling tasks, relational domains still lack a foundation model that transfers across heterogeneous schemas, graph structures, and functional dependencies. Relational Transformers [3] show promising zero-shot performance, but still fall short of the broad out-of-distribution robustness exhibited by large language models. At Stanford, I am working with Prof. Jure Leskovec on extending the model to few-shot long-context settings, which will allow the model to observe functional dependencies directly in context and thereby achieve stronger generalization on out-of-distribution data. This direction requires addressing substantial systems-level challenges: scaling to long contexts demands making attention masks block sparse to mitigate the quadratic slow-down of standard attention. At the same time, the modeling side requires constructing meaningful contexts. Specifically, retrieving examples that are most relevant to the target relational query to maximize the utility of each few-shot demonstration. Through this work, I have learned to reason across both systems and modeling constraints, iterate on architectural choices, and perform systematic empirical experiments to guide exploration.

Beyond my work in machine learning systems and modeling, my earlier projects deepened my appreciation for theoretical structure and rigorous reasoning. With Prof. Swaprava Nath at IIT Bombay, I worked on balanced group partitions and proved the existence of envy-freeness for general graphs [2], a result published at the *International Conference on Autonomous Agents and Multiagent Systems*. A key step in our proof required identifying an obscure lemma guaranteeing existence of st-numbering for biconnected graphs. By decomposing general graphs into blocks of biconnected components and applying this property, we could construct the desired partition. Arriving at this insight required deep literature search, extensive discussions, and iterative refinement of ideas with co-authors. This project taught me how collaborative reasoning and persistence come together in solving a challenging theoretical problem. I also worked with Prof. Ashish Goel on designing and implementing an algorithm to improve expert-question selection during large-scale deliberative processes, building on ideas from justified representation to ensure that participants interest were fairly captured. This experience helped me bridge theoretical ideas with practical system design and strengthened my interest in research that integrates principled reasoning with real-world impact.

At Stanford, I hope to continue working with Prof. Jure Leskovec, building on recent advances in graph and relational transformers and enabling models to generalize across heterogeneous relational domains and unseen schemas. With Prof. Stefano Ermon, I am interested in extending diffusion and energy-based generative models to structured settings, designing models capable of generating or completing relational configurations under distribution shift by leveraging principles from probabilistic inference and combinatorial reasoning. With Prof. Chris Ré, I would be excited to work on automated supervision and data-centric learning pipelines—using programmatic labeling, weak supervision, and large-scale data systems to build models whose internal representations and reasoning behaviors are grounded in higher-quality, systematically curated training signals. These directions align closely with my goal of understanding how models internalize structure and how principled design choices can produce more reliable, general-purpose reasoning systems.

Pursuing a PhD would place me in the kind of continuously progressing academic environment where I can keep learning new ideas, engage with challenging and impactful problems, and take full ownership of my research directions. Looking ahead, I hope to deepen my understanding of the foundations, capabilities, and limitations of modern learning algorithms and use these insights to develop models and algorithms that reason more reliably and generalize more effectively. My experiences at IIT Bombay and Stanford have prepared me for this path by giving me a foundation that spans theoretical analysis, system building, and empirical evaluation. A PhD at Stanford would place me in an environment where foundational ideas are developed from first principles and tested in the systems that bring them to life, enabling the long-term exploration required to pursue these questions meaningfully. I look forward to growing within a community that values deliberate inquiry and that supports long-term scientific progress.

References

[1] Harshvardhan Agarwal and Sunita Sarawagi. "The Missing Alignment Link of In-Context Learning on Sequences". In: *Proceedings of the Forty-Second International Conference on Machine Learning*. 2025.

[2] Pulkit Agarwal et al. "Harmonious Balanced Partitioning of a Network of Agents". In: *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*. 2025.

[3] Rishabh Ranjan et al. *Relational Transformer: Toward Zero-Shot Foundation Models for Relational Data*. 2025. arXiv: 2510.06377 [cs.LG].