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 to adapt them when the underlying data is not naturally sequential. 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 architectural modifications, informed by an understanding of model internals, can lead to more robust and 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. We built a controlled synthetic framework to isolate these effects and demonstrated that language models often fail to recover the alignment in-context. This research resulted in our [1] paper at the *International Conference on Machine Learning*; gave me a mechanistic perspective on when and why transformers succeed or fail at ICL. It also strengthened my interest in grounding model design in a deeper understanding of the computations transformers perform internally.

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 [4] 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, I worked on balanced group partitions and proved the existence of envy-freeness for general graphs [2]. 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 the Stanford Online Deliberation Platform, where we integrated language-embedding models to improve expert-question selection during large-scale deliberative processes [3].

At Stanford as a PhD, 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 allow me to engage more fully with the parts of research I enjoy most: reading deeply, identifying gaps in existing methods, and collaborating to develop ideas that push our understanding forward. My experiences at IIT Bombay and Stanford have shaped the way I approach research: grounding ambitious ideas in careful reasoning, building systems that embody these ideas, and validating them through thoughtful empirical work. Going forward, I seek to study the capabilities, and limitations of modern learning algorithms both theoretically and empirically, with the long-term goal of contributing to models that reason reliably across diverse forms of structure. With its deep-rooted academic traditions, collaborative research culture, and world-leading faculty, Stanford is the ideal environment for me to develop as a researcher. Through a PhD, I hope to gain the depth, breadth, and intellectual community necessary to work toward the long-term scientific challenge that inspires me most; understanding and ultimately helping to solve intelligence.

add concrete examples to first paragraph. increase paragraphs lengths fix second last paragraph fix last paragraph

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] Soham De et al. *Question the Questions: Auditing Representation in Online Deliberative Processes*. 2025. arXiv: 2511.04588 [cs.AI].

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