Statement of Purpose

Jian Vora *Ph.D. Applicant*

I am interested in building trustworthy autonomous agents for decision-making. Particularly, I am excited about i) AI for Science (using agents to accelerate scientific discovery) and ii) the Science of AI (interpretability and robustness from a sparsity standpoint). I have been fortunate enough to have a diverse and fulfilling set of research experiences which have convinced me that grad school would be the ideal next step for me. In the subsequent sections, I shall focus on relevant research experiences followed by what I would like to explore in grad school along that line of research.

AI Agents for Scientific Discovery: Disruptive ideas in science have reduced in many fields when compared to the last century. Can we have (LLM-based) AI agents that perform the end-to-end scientific experimentation cycles for us – perform a literature review, generate a hypothesis, design experiments, and finally test the hypothesis to push the boundaries of science? I worked with Jure Leskovec and Percy Liang on developing agents for:

- **Machine Learning**: The agent was given system access (read/write files, execute code, etc.) and was asked to solve machine learning tasks provided an initial prompt and a working directory with some initial code and data. We developed a framework to evaluate agents for both their success rates, and efficiency (tokens used), and characterized common failure modes such as hallucination and lack of long-term planning. This work has been accepted at the NeurIPS 2023 FMDM Workshop and is under review ICLR 2024.
- **Biology**: The agent was asked to perform CRISPR perturbation experiments to identify genes from the entire human genome that influence the production of a given target protein. LM Agents were able to navigate the large search space efficiently with their useful priors and came up with very interpretable pathways, better hit rates compared to Bayesian experimental design algorithms, and robustness to noisy readouts typical of biological experiments. This work has excited biologists at the Gladstone Institute (UCSF) and we are collaborating with them to test the agent on data from actual lab trials.

The above experiences showed me both the promises and shortcomings of current language-based agents such as:

- **Lack of Scientific Creativity**: The agent performed surprisingly well in experimental design and execution but lacked the creativity to come up with interesting hypotheses. I would like to push scientific reasoning and creativity into these agents which could significantly change the way scientific investigations are made.
- **Designing Memory for Agents**: Current transformer-based models are stateless, limited by their context lengths. Interestingly, we observed that maintaining a simple research log and retrieving from it, to be passed in the prompt, degraded the performance of the ML agent. I am interested in exploring this further to design a hierarchical (multimodal) memory module (unlike a vector database) allowing for efficient retrieval for AI agents which has a range of applications not just for decision-making but also long-form content generation.
- **Internal World Model vs External Feedback**: An issue with the current agents is that is hard to incorporate observations in decision-making to override the agents' prior knowledge. This looks like a fundamental limitation for having self-improving agents. Exploring the interplay between external observations/memory and the agents' internal world model (here, an LLM) seems crucial. I would like to push for self-improving AI agents without the need for explicit rewards or demonstrations.

Efficient and Reliable Machine Learning: Current models keep on getting bigger which obscures interpretability, robustness, and has larger training and inference system costs. On the other hand, it has been shown that representations learned by these models lie in a low-dimensional space and models can be pruned up to a significant extent without affecting performance. I would like the explore whether can we leverage the fact that most real-world data lies on a low-dimensional manifold and train sparse models incorporating this inductive bias. Sparse models come with the added advantage of easier mechanistic interpretability which is useful when these models are used for high-stakes applications.

- **Sparse Models**: I have worked on sparse models in my undergrad that leverage some useful prior about the underlying data distribution. I worked with Ajit Rajwade on compressed sensing recovery under sensing matrix perturbations (ICASSP '21), joint density estimation from marginals using low-rank tensor factorization (IEEE SSP '21), proving that random projections of data from a mixture of log-concave densities are provably distributed as a gaussian mixture on the subspace (Technical Report).

- **Trustworthy Machine Learning**: I have worked independently on trying to understand the adversarial robustness of current models better. For graphs, I showed that simply performing GNN inference on k-hop ego nets instead of the entire graph improves robustness across many common adversarial attacks. This has been accepted as a **solo author** paper at the NeurIPS 2023 GLFrontiers Workshop. In another work, we showed that the l_1 norm of LIME weights is a good scoring function for the adversarial robustness of black-box models and showed that more robust models have a small score (which meant sharper explanations). This has been accepted at the ICML 2022 AdvML Workshop. The above experiences of researching independently taught me how to identify interesting problems, carry out the entire experimentation cycle, and write up the results while being self-motivated along the way. I liked picking up problems that were wide enough in relevance and coming up with simple yet useful insights for those.

  Currently, I am working with Prof. Fei-Fei Li on post-processing policies to reduce hallucination in the robotics domain by incorporating the physical world knowledge in the language model. We are trying to add a control vector to the language model used as a policy to ground it to the planning domain and experimenting on VirtualHome and Behavior environments.

I would like to continue working on efficient and reliable decision-making which is extremely crucial when these agents are deployed in the wild.

Industry Research Experience and Impact: I have some industry research experience working on real-world data and large-scale models. I worked at Twitch on sequential ranking in recommendations (a huge +9% relative improvement in offline NDCG), at AWS AI on pretraining a multimodal speech-text foundation model (currently used by the AWS Lex ASR Service), and at Hitachi Japan on activity detection in videos (landed 3rd in the Trecvid competition). I worked with Stanford Design School to design a food recommender system that takes into account the long-term health of the users in addition to their interaction. This was posed as a multi-objective (CVaR) policy optimization problem and the model was deployed at the Stanford dining halls. We showed the applications of our PAC mode estimation work (AISTATS '22) in the context of Indian election polls and verification in blockchains. I believe skills related to engineering and applied research would be invaluable in my grad school experience.

Having seen the rewards and struggles of research, I am convinced that a Ph.D. would be a rewarding time for me. I was fortunate enough to work on a wide range of machine learning problems and I am excited about the future, particularly when we are seeing the unification of various fields of machine learning. I enjoyed TAing many AI courses at Stanford (CS234, CS221, CS224V, CS236G) and mentoring students, both of which would help in making my grad school experience fruitful. Post my graduate studies, I would like to pursue a research career either in academia or industry.

References