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 AI agents to accelerate scientific discovery) and ii) the Science of AI (interpretability and robustness). 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: The number of disruptive ideas in science has reduced in many fields when compared to the last century [1]. Can we have 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 provided tools (read/write files, execute code, etc.) and was asked to solve machine learning tasks provided an initial prompt and a working directory. We developed a framework to evaluate agents, designed our LM based agent, and characterized common failure modes. This work has been accepted as an Oral at the NeurIPS 2023 FMDM Workshop and is under review at ICLR 2024 [2].
- **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, came up with very interpretable pathways, and queried literature to design the next batch [3]. This work has excited biologists at the Gladstone Institute 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 testable hypotheses. I would like to push for scientific creativity into these agents which could significantly change the way scientific investigations are made. This requires studying whether these agents can actually reason/self-critique or are performing approximate retrieval. I am interested in furthering system-2 [4] reasoning and planning capabilities of these agents.
- **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 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 which are hard to obtain in many real-world applications and can hinder scalability.
- **Interpretability and Reliability**: Hallucinations, along with a lack of interpretability and susceptibility to adversarial inputs, are common challenges that impede the widespread adoption of language models in the realm of scientific discovery. I am currently working with Prof. Fei-Fei Li on post-processing LM policies to reduce hallucination in the robotics domain by grounding it in the planning domain. In the subsequent sections, I shall describe some of my work on robustness and foundations of machine learning.

Science of AI: I am interested in trying to understand the current machine learning models better, which is especially important for applications in diverse domains. Some of my previous works have delved into making current machine learning models more robust and both sample and compute efficient.

- **Robust Machine Learning**: In a course, we explored how models are prone to adversarial attacks. This sparked my interest, leading to my independent research on enhancing model robustness. I found that using GNN inference on k-hop ego networks instead of whole graphs increased resistance to attacks, a finding I

published as a solo author at NeurIPS 2023 GLFrontiers Workshop [5]. In another study, accepted at the ICML 2023 AdvML Workshop [6], we showed that the l_1 norm of LIME weights could effectively gauge a model's robustness, with more robust models providing clearer explanations. 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.

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 to 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.

- **Machine Learning Foundations**: 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 [7]), joint density estimation from marginals using low-rank tensor factorization (IEEE SSP '21 [8]), 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 [9]), learning multimodal distributions tractably on latent spaces (Technical Report [10]).

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 +9% relative improvement in offline NDCG, AMLC '23 [11]), 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. 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 is running at the Stanford dining halls. We showed the applications of our PAC mode estimation work (AISTATS '22 [12]) 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 graduate 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 multiple AI courses at Stanford 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

[1] Leahey E. Funk R.J. Park, M. Papers and patents are becoming less disruptive over time. *Nature 613, 138144 (2023). https://doi.org/10.1038/s41586-022-05543-x*, 2023.

[2] Qian Huang, **Jian Vora**, Percy Liang, and Jure Leskovec. Benchmarking large language models as AI research agents. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

[3] Yusuf Roohani, **Jian Vora**, Qian Huang, Percy Liang, and Jure Leskovec. Ai guided crispr perturbation experiments. *Under Preparation*, 2023.

[4] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[5] **Jian Vora**. Gnn predictions on k-hop egonets boosts adversarial robustness. In *NeurIPS 2023 Workshop, New Frontiers in Graph Learning*, 2023.

[6] **Jian Vora** and Pranay Reddy Samala. Scoring black-box models for adversarial robustness. In *The Second Workshop on New Frontiers in Adversarial Machine Learning, ICML*, 2023.

[7] **Jian Vora** and Ajit Rajwade. Compressive signal recovery under sensing matrix errors combined with unknown measurement gains. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5105–5109, 2021.

[8] **Jian Vora**, Karthik S. Gurumoorthy, and Ajit Rajwade. Recovery of joint probability distribution from one-way marginals: Low rank tensors and random projections. In *2021 IEEE Statistical Signal Processing (SSP)*, 2021.

[9] **Jian Vora** and Vivek Borkar. Multimodal density estimation from random linear compressive projections. *Report Link*, 2020.

[10] **Jian Vora**, Isabel Valera, Guy Van den Broeck, and Antonio Vergari. Plug&play tractable multimodal probabilistic learning, 2022. *to be submitted to Transactions on Machine Learning (TMLR)*,.

[11] **Jian Vora**, Edgar Chen, Nikita Mishra, and Saad Ali. Sequential consumption-aware ranking model for recommendations at twitch. *AMLC Workshop on Personalization and Ranking*, 2023.

[12] Shubham Anand Jain, Rohan Shah, Sanit Gupta*, Denil Mehta*, Inderjeet J Nair*, **Jian Vora***, Sushil Khyalia, Sourav Das, Vinay J Ribeiro, and Shivaram Kalyanakrishnan. Pac mode estimation using ppr martingale confidence sequences. In *International Conference on Artificial Intelligence and Statistics*, pages 5815–5852. PMLR, 2022.